(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification⁷: C12N 15/11, C12Q 1/68, G01N 33/53

(21) International Application Number: PCT/US01/48527

(22) International Filing Date:
14 December 2001 (14.12.2001)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **ROSETTA INPHARMATICS, INC.** [US/US]; 12040 115th Avenue, N.E., Kirkland, WA 98034 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*:   **MARTON, Matthew** [US/US]; 2415 NW 91st, Seattle, WA 98117 (US). **JONES, Allan** [US/US]; 11410 NE 124th Street #454, Kirkland, WA 98034 (US).

(74) **Agents: ANTLER, Adriane, M.** et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) **Designated States** *(national)*: CA, JP, US.

(84) **Designated States** *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

(54) **Title:** SAMPLE TRACKING USING MOLECULAR BARCODES

(57) **Abstract:** The present invention relates to methods and compositions for providing internal identification in samples. In particular, the invention relates to methods of tracking samples using a molecular barcode made up of spike-in tags added to the sample before or during processing. The spike-in tags can be co-processed with the sample through a plurality of processing steps such that any transfer of samples to different containers or solid phases during processing also transfers the molecular barcode. Samples made up of any macromolecule (including DNA, RNA, and proteins) may be tracked. The invention further relates to a method of using a computer to determine the identity of a particular sample by comparing the profile of spike-in tags which make up the molecular barcode of the sample with a database of sample identities associated with the particular molecular barcode added. Methods of the invention can also be used in quality control applications.

SAMPLE TRACKING USING MOLECULAR BARCODES

1. FIELD OF THE INVENTION

The present invention relates to methods and compositions for providing
internal identification in samples. In particular, the invention relates to methods of tracking
samples using a molecular barcode made up of spike-in tags added to the sample before or
during processing. In a preferred embodiment, this invention also relates to the
unambiguous labeling of samples which are to be subjected to high throughput processing.
The invention further relates to a method of using a computer to determine the identity of a
particular sample by comparing the profile of spike-in tags which make up the molecular
barcode of the sample with a database of sample identities associated with the particular
molecular barcode added. Methods of the invention can also be used in quality control
applications.

2. BACKGROUND OF THE INVENTION

In complex processes involving considerable human intervention, methods
are needed to ensure that samples are handled and identified correctly. It has been
demonstrated that samples are at considerable risk for being inadvertently and unknowingly
mixed up. For example, handling errors occurred in the sample processing during the
sequencing of the *Drosophila* genome
(http://www.discovery.com/news/briefs/20000420/tech_genome.html). Some sequences
reported to be from the fly were perfect matches to human sequences (Adams et al., 2000,
The genome sequence of *Drosophila melanogaster*, *Science* 287:2185-95). This suggests
that samples were either mixed up in the laboratory or contaminated with human samples.
When this happens, the resulting data from the sample processing is flawed and thus can be
rendered useless. Handling errors also can occur during processing of clinical samples.
Donated blood is typically screened for viral markers to reduce the risk of virus
transmission by transfusion. False negative testing errors were examined and some sample
mix-ups and sample processing errors were found to be present (Busch et al., 2000,
Transfusion 40:585-589).

In general, many sample processing protocols involve the transfer of
materials from the container in which the sample was collected to at least one other
container. Thus, even where the original sample container has been correctly labeled, there
is a risk that samples taken from different sources will be mixed up during the subsequent
transfers for analysis. This problem is particularly applicable to processes that involve

numerous transfers throughout the protocol. For example, microarray technology is one such assay. Although various formats of microarrays are currently used, all DNA array technologies employ nucleic acid probes, (*i.e.*, nucleic acid molecules having defined sequences) to selectively hybridize to, and thereby identifying and measuring the

5   abundances of or relative abundances of, complementary nucleic acid sequences in a sample. In these technologies, a set of nucleic acid probes, each of which has a defined sequence, is immobilized on a solid support in such a manner that each different probe is immobilized to a predetermined region. The set of immobilized probes or the array of immobilized probes is contacted with a sample containing labeled nucleic acid species so

10 that nucleic acids having sequences complementary to an immobilized probe hybridize or bind to the probe. After separation of, *e.g.*, by washing off, any unbound material, the bound, labeled sequences are detected and measured. The amount of labeled sequence hybridized to each probe in the array is used as a measure of the abundance of the sequence species in the sample (*see, e.g.*, Schena et al., 1995, *Science 270*:467-470; Lockhart et al.,

15 1996, *Nature Biotechnology 14*:1675-1680; Blanchard et al., 1996, *Nature Biotechnology 14*:1649). Using DNA array expression assays, complex mixtures of labeled nucleic acids, *e.g.*, mRNA molecules or nucleic acids derived from mRNA molecules from a cell or a population of cells, can be analyzed.

DNA array technologies have made it possible, *inter alia*, to monitor the

20 expression levels of a large number of genetic transcripts at any one time (see, *e.g.*, Schena et al., 1995, *Science 270*:467-470; Lockhart et al., 1996, *Nature Biotechnology 14*:1675-1680; Blanchard et al., 1996, *Nature Biotechnology 14*:1649; Shoemaker et al., U.S. Patent Application Serial No.09/724,538, filed on November 28, 2000). DNA array technologies have also found applications in gene discovery, e.g., in identification of exon structures of

25 genes (see, *e.g.*, Shoemaker et al., U.S. Patent Application Serial No. 09/724,538, filed on November 28, 2000).

By simultaneously monitoring tens of thousands of genes, microarray technologies have allowed, *inter alia*, genome-wide analysis of mRNA expression in a cell or a cell type or any biological sample. Aided by sophisticated data management and

30 analysis methodologies, the transcriptional state of a cell or cell type as well as changes of the transcriptional state in response to external perturbations, including but not limited to drug perturbations, can be characterized on the mRNA level (see, *e.g.*, U.S. Patent No. 6,203,987; Stoughton et al., U.S. Patent Application Serial No. 09/220,275 (filed on December 23, 1998); Stoughton et al., International Publication No. WO 00/39336

35 (published July 6, 2000); Friend et al., International Publication No. WO 00/24936

(published May 4, 2000)). Applications of such technologies include, for example, identification of genes which are up regulated or down regulated in various physiological states, particularly diseased states. Additional exemplary uses for DNA arrays include the analyses of members of signaling pathways, and the identification of targets for various

5   drugs. See, *e.g.*, Friend and Hartwell, International Publication No. WO 98/38329 (published September 3, 1998); Friend and Stoughton, International Publication No. WO 99/59037 (published November 18, 1999); U.S. Patent Nos. 6,132,969; 5,965,352; 6,218,122.

While methods for the design, collection and analysis of gene expression

10   data using microarrays has been described (see U.S. Patent Nos. 5,510,270; 5,556,752; 5,578,832; 6,028,189; 6,171,794; and 6,228,593), what is needed is a system and method suitable for collecting and storing large quantities of microarray information without the possibility of accidental misidentification of sample and corresponding data.

In a clinical setting, large numbers of samples must be collected and

15   processed. Information about the sample donors must be maintained throughout the processes of sample preparation, hybridization, data collection and analysis to facilitate subsequent diagnoses. The quantity of information to store and correlate is vast and can be compounded by information management equipment and other laboratory resources that may be shared with other projects. A single laboratory may service many clients, each

20   client in turn requesting completion of multiple projects. Mishandling or sample misidentification mistakes could be of great harm when samples are used in the diagnosis of disease states and drug response. There are documented cases that this indeed is a problem. For example, Ladenson (1975, *Clin. Chem.* 21:1648-1653) followed test results performed by a hospital's clinical chemistry laboratory over a 22 month period and found that

25   specimen misidentification was routinely occurring.

This application describes a method designed to prevent such mislabeling and mishandling mistakes by the introduction of molecules, termed spike-in tags, that are added to the sample and carried through the purification and data analysis process. The profile of which particular spike-in tags have been added to a sample are referred to as the

30   sample's molecular barcode. Since the spike-in tags are added to the original sample or to the container which will hold the original sample, there is no chance that mishandling or mislabeling will lead to sample misidentification errors.

The herein disclosed invention describes methods for the design and production of spike-in tags, methods to impart a molecular barcode to a sample by the

35   addition of a particular profile of spike-in tags, and methods for analysis of the information

obtained from sample processing that permit the determination of the molecular barcode of the sample. This invention also includes a relational database that contains information concerning each molecular barcode, such as which profile of spike-in tags make up the molecular barcode and the identity of which sample that particular molecular barcode was

5   added.

Citation or discussion of a reference herein shall not be construed as an admission that such is prior art to the present invention.

3. SUMMARY OF THE INVENTION

10   The present invention relates to methods and compositions for providing internal identification in samples. In particular, the invention relates to methods of tracking samples using a molecular barcode made up of spike-in tags added to the sample. The particular combination of spike-in tags added to a sample is what gives the sample its unique identification termed its molecular barcode. The spike-in tags can be added to the

15   sample at various times during the collection or processing of a sample. In a preferred embodiment, the spike-in tags are added to a sample before sample processing. In a more preferred embodiment, the spike-in tags are added to a sample at the time of collection.

The methods of the invention are particularly applicable to samples subjected to complex multi-step assays. Assays where numerous sample transfers are performed

20   increases the risk for improper handling or labeling (*e.g.*, in an environment where many samples are being processed simultaneously). A preferred embodiment of the invention is for sample tracking during microarray assays. In a more preferred embodiment the microarray is an oligonucleotide array. In a most preferred embodiment, the oligonucleotide array is used to monitor the levels of a large number of RNA transcripts for expression

25   analysis.

In one embodiment, the present invention is directed to a method for providing internal identification of a sample comprising RNA or protein from a biological specimen such that said sample can be tracked through a processing step, comprising the following steps in the order stated:

30   (a)   adding one or more spike-in tags that are RNA or protein
                   molecules that form a first molecular barcode to said sample;
            (b)   processing said sample through a plurality of processing
                   steps;
            (c)   determining the identity of the spike-in tags in said processed

35   sample to determine a second molecular barcode; and

(d)   comparing said first molecular barcode with said second
        molecular barcode to determine whether said first molecular
        barcode does not differ from said second molecular barcode,
        wherein a match indicates that said sample has been tracked

5   through said processing step.

        In another embodiment, the present invention is directed to a method for
tracking a plurality of samples, each sample comprising RNA from a different biological
specimen, or nucleic acid derived therefrom, comprising the following steps in the order
stated:

10   a)   adding to each sample one or more RNA molecules that are spike-in tags
                such that the one or more spike-in tags in each sample form a molecular
                barcode that distinguishes each sample from said other samples;

        b)   processing each of said samples wherein said processing comprises
                transferring RNA from each sample, or nucleic acid derived therefrom, to a

15   different container or solid phase;

        c)   determining the identity of the spike-in tags in each processed sample to
                determine the molecular barcode in each processed sample; and

        d)   comparing the determined molecular barcodes for at least some of said
                samples to a compendium of molecular barcodes associated with sample

20   identity to determine the identities of said at least some samples.

        In another embodiment, the present invention is directed to a method for
tracking a plurality of samples, each comprising RNA from a different biological specimen,
or nucleic acid derived therefrom, comprising the following steps in the order stated:

        a)   adding a molecular barcode to each sample that distinguishes each sample

25   from said other samples, said molecular barcode composed of one or more
                RNA molecules that are spike-in tags, said spike-in tags each consisting of
                an RNA molecule comprising a sequence not known (i) to be present in any
                naturally occurring nucleic acid, or (ii) to encode any naturally occurring
                protein;

30   b)   processing each of said samples, wherein said processing comprises
                contacting a positionally-addressable array of polynucleotide probes with
                RNA molecules from said sample, or nucleic acid derived therefrom, under
                conditions conducive to hybridization between said probes and said RNA
                molecules or nucleic acid derived therefrom, wherein said array comprises a

35   plurality of polynucleotide probes of different nucleotide sequences bound to

different regions of a support, wherein said plurality of polynucleotide probes comprises (i) probes hybridizable to said spike-in tags, and (ii) probes hybridizable to said RNA molecules or nucleic acid derived therefrom; and detecting or measuring hybridization between said polynucleotide probes and (i) said spike-in tags, and (ii) said RNA molecules or nucleic acid derived therefrom;

c)   determining the identity of the spike-in tags in each processed sample to determine the molecular barcode in each processed sample; and

d)   comparing the determined molecular barcodes for at least some of said processed samples to a compendium of molecular barcodes associated with sample identity to determine the identities of said at least some samples.

In another embodiment, the present invention is directed to a method for providing internal identification of a sample comprising RNA from a biological specimen such that said sample can be tracked through a processing step comprising the following steps in the order stated:

(a)   processing said sample wherein said processing comprises making cDNA from said RNA in said sample, wherein subsequent or during said making one or more spike-in tags that are isolated DNA molecules that form a first molecular barcode are added to said processed sample at a time when the nucleic acids derived from the sample RNA are cDNA, wherein said DNA molecules each comprise sequences not known (i) to be present in any naturally occurring nucleic acid, or (ii) to encode any naturally occurring protein;

(b)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(c)   comparing said first molecular barcode with said second molecular barcode, wherein a match indicates that said sample has been accurately tracked.

In another embodiment, the present invention is directed to a method of providing internal identification of a sample from a biological specimen such that said sample can be tracked through a possessing step comprising the following steps in the order stated:

(a)   adding one or more isolated spike-in tags that form a first molecular barcode to said sample, wherein said spike-in tags are biopolymers; wherein when said biopolymers are DNA

molecules, said DNA molecules comprise a sequence not known (i) to be present in any naturally occurring nucleic acid, or (ii) to encode any naturally occurring protein;

(b)   processing said sample through a plurality of processing steps wherein said processing comprises contacting a positionally-addressable array of probes with biopolymers from said sample, or biopolymers derived therefrom, under conditions conducive to interaction between said probes and said sample biopolymer molecules or biopolymers derived therefrom, wherein said array comprises a plurality of probes of different sequences bound to different regions of a support, wherein said plurality of probes comprises (i) probes capable of interacting with said spike-in tags, and (ii) probes capable of interacting with sample biopolymers or biopolymers derived therefrom; and detecting or measuring interaction between said probes and (i) said spike-in tags, and (ii) said sample biopolymers or biopolymers derived therefrom;

(c)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(d)   comparing said first molecular barcode with said second molecular barcode to determine whether said first molecular barcode does not differ from said second molecular barcode, wherein a match indicates that said sample has been tracked through said processing step.

In another embodiment, the present invention is directed to a method for providing internal identification of a sample comprising protein from a biological specimen such that said sample can be tracked through a processing step comprising the following steps in the order stated:

(a)   adding one or more spike-in tags that are protein molecules that form a first molecular barcode to said sample;

(b)   processing said sample;

(c)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(d)   comparing said first molecular barcode with said second molecular barcode, wherein a match indicates that said

sample has been accurately tracked through said processing step.

In another embodiment, the present invention is directed to an RNA molecule consisting of:

5   (a)   a sequence in the range of 20-5000 contiguous nucleotides of which at least 50% are random; and

(b)   a poly A tail.

In another embodiment, the present invention is directed to an RNA molecule comprising:

10   (a)   a sequence not known (i) to be present in any naturally occurring nucleic acid, or (ii) to encode any naturally occurring protein; and

(b)   a poly A tail.

In another embodiment, the present invention is directed to a method of using a computer to determine the identity of a processed sample containing a molecular

15 barcode wherein said molecular barcode comprises one or more spike-in tags, comprising the following computer-implemented steps:

(a)   receiving a first data structure comprising the molecular barcode of said processed sample; and

(b)   comparing said first data structure to a plurality of data structures in a

20   database, each data structure comprising a molecular barcode associated with sample identity, to identify a molecular barcode that is a match, wherein said match indicates the identity of said processed sample.

In another embodiment, the present invention is directed to a computer system for determining the identity of a sample, said computer system comprising:

25   one or more processor units; and

one or more memory units connected to said one or more processor units,

said one or more memory units containing one or more programs which cause said one or more processor units to execute steps of:

(a)   receiving a first data structure comprising the molecular barcode of a

30   processed sample; and

(b)   comparing said first data structure to a plurality of data structures in a database, each data structure comprising a molecular barcode associated with sample identity, to identify a molecular barcode that is a match, wherein said match indicates the identity of said processed sample.

35

In another embodiment, the present invention is directed to computer readable medium containing an encoded data structure, said data structure comprising:

    (a)   a digital representation of the identity of each sample associated with a molecular barcode;

5   (b)   a digital representation of the particular spike-in tags comprising each molecular barcode; and

    (c)   a digital representation of which of said molecular barcodes has been added to which of said samples.

10
3.1 DEFINITIONS

As used herein, the term "spike-in tag" is a molecule that can be added to a biological specimen or a sample to serve as part of a molecular barcode for that sample.

As used herein, the term "molecular barcode" is a sample identifier provided by one or more spike-in tags. Samples are assigned a unique molecular barcode which

15 allows for sample tracking throughout multi-step processing.

As used herein, the term "biological specimen" is broadly defined to include any cell, tissue, organ or multicellular organism. A biological sample is derived from a biological specimen, for example, from cell or tissue cultures *in vitro*. Alternatively, a biological sample can be derived from a living organism or from a population of single cell

20 organisms. A biological sample can be composed of one or more different biopolymers, *e.g.*, genomic DNA, RNA, or proteins.

As used herein, the terms "epitope" or "antigenic determinant" is a specific portion of a macromolecule that can be bound by an antibody. In polypeptides, linear epitopes are epitopes formed by a contiguous segment of the primary amino acid sequence.

25 Conformational epitopes are formed by amino acid residues from separate portions of the linear polypeptide that are spatially juxtaposed upon folding.

4. BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates by way of example the synthesis of synthetic RNA

30 molecules for use as spike-in tags. The plasmids were constructed with common molecular biological techniques known in the art. First, oligonucleotides that were 60 nucleotides in length were synthesized with the middle 52 nucleotides having a random sequence. These oligonucleotides were cloned into pSP64-E1A plasmids that had been previously digested to completion with XbaI and BamHI. The plasmids were amplified, isolated and

35 sequenced. Those clones containing oligonucleotide inserts were linearized with EcoRI and

purified before serving as a template for *in vitro* transcription with SP6 DNA polymerase. Resulting RNA had a poly A tail from the plasmid. Spike-in tag RNA was purified with RNeasy columns before use.

Figures 2A-B illustrate the processing that can be involved in preparing a biological sample for data analysis using an oligonucleotide microarray. (A) Method of production of cRNA from sample RNA is depicted. Nucleic acid spike-in tags can be added at any step in the process provided that the spike-in tags are of the same type of biopolymers as in the sample at the time of introduction of the spike-in tags and/or are being assayed in the multi-step process. (B) A twenty-step process that can be used to prepare a biological sample of cells for examination by oligonucleotide microarray hybridization. This process involves multiple tube transfers, any one of which is subject to mishandling or mislabeling.

Figures 3A-B illustrate that spike-in tag RNA can be added to total RNA and be amplified, labeled, hybridized, and detected on the ink jet oligonucleotide microarray. (A) microarray image after cRNA hybridization. (B) ratios of the spike-in probes detected in the samples.

Figures 4A-C illustrate microarray images from three experiments in which spike-in tag RNA molecules were detected on microarrays while in the presence of other sample RNA molecules. (A) cDNA was made using an oligo dT primer to reverse transcribe total sample RNA. (B) cRNA was made from total sample RNA by reverse transcription before being transcribed *in vitro*. (C) cRNA was made from total sample RNA using the protocol outlined in Figure 2A. In all three experiments, spike-in tags were detected on the microarrays.

Figures 5A-C illustrate that spike-in tag RNA molecules do not significantly impact detection of cRNA molecules originating from sample RNA. (A) plot showing the ratio of intensities detected at each probe on the microarray. (B) plot of the reverse fluor pair of 5A. (C) the correlation of the ratios in Figures 5A and 5B.

Figures 6A-B illustrate two different identification schemes used to track a sample in an individual well of a 96-well plate. The scheme in (A) is based on the position of the well in the grid of the 96-well plate whereby each row and column is assigned a unique set of spike-in tag RNA (upper panel). The lower panel is an example of three different molecular barcodes that can be given to three different samples The scheme in (B) is based on a sequential numbering of samples whereby each digit of the identification number is represented by a different spike-in tag RNA. The upper panel is an example of a simple key where each digit is assigned a unique spike-in tag. The lower panel is an

example of four different molecular barcodes that can be given to four different samples. In this example, three spike in tags are added to each sample to make the molecular barcode.

Figure 7 schematically illustrates the use of molecular barcodes in data quality control. Prior to sample processing, an association is made between information

5   describing a particular sample and the unique molecular barcode assigned to that sample. After sample processing, the actual molecular barcode present in that particular sample is deduced and is associated with the data. The expected and actual molecular barcodes are compared. The data meets the quality standards only if there is a match between the molecular barcodes.

10

5. DETAILED DESCRIPTION OF THE INVENTION

The object of the present invention is to unambiguously label a biological sample to be processed with a unique internal identifier. This object is fulfilled by the addition of a molecular barcode to the biological sample composed of at least one, but more

15   preferably more than one, spike-in tags. The spike-in tags are processed in the same manner as, and in conjunction with, the biological sample. Upon completion of the processing, each spike-in tag is identified and contributes to the unique molecular barcode. The molecular barcode given to the biological sample at the time of spike-in addition can then be compared with data resulting from processing to confirm sample identity.

20   The sample which is tracked according to the invention comprises macromolecules (e.g., biopolymers) from or derived from a biological specimen (e.g., cell, tissue, or organism). In a specific embodiment, a plurality of samples from different biological specimens are tracked, e.g., through concurrent processing steps. For example, the different biological specimens can be different individuals, organisms of different or the

25   same species, different cell types (e.g., of different tissue origin or different developmental stage of the same or different species or organisms), cells or organisms subjected to different perturbations or no perturbations (perturbations being an exogenous effect, e.g., exposure to drug, disease, environmental change, etc.); autologous, allogeneic, or xenogeneic, etc.

30

5.1 SPIKE-IN TAGS

Spike-in tags are molecules added to samples that do not interfere with the results of the intended sample processing. In a preferred embodiment, the spike-in tags isolated or purified from other like biopolymers. Samples can be processed in any number

35   of ways. Preferably, the molecular nature of the spike-in tag is compatible with the type of

processing to be performed to allow for co-processing of the spike-in tags with the sample. In one embodiment, gene expression is to be assayed thus the spike-in tags are preferably mRNA molecules. In another embodiment, protein expression or activity is to be assayed thus the spike-in tags are preferably peptides. In another embodiment, an aneuploidy (see,

5   *e.g.* U.S. Provisional Patent Application No. 60/250,597 filed December 1, 2000) or a single nucleotide polymorphism is to be detected thus the spike-in tags are preferably DNA. In other embodiments, the spike-in tags can be different biopolymers than are present in the sample, *e.g.*, nucleic acid spike-in tags used to track polypeptides.

According to the present invention, in a specific embodiment, sample

10   tracking is the process of adding spike-in tags (to create a molecular bar code) to a sample at the beginning of a multi-step process, performing the process, and determining the identity of the spike-in tags in the sample to read the identity of the molecular barcode after the process. One use of sample tracking is quality control. If the barcode at the end of a process does not match the barcode added at the beginning of the process, the data is

15   disregarded. Another use of sample tracking is sample identification. At the end of a process, the identities of the barcodes can be ascertained. This information can be used to match a particular profile of spike-in tags added to a particular sample to create the barcode, to then identify the sample.

20   5.1.1 <u>DESIGN OF SPIKE-IN TAG MOLECULES</u>

In the method of the present invention, the collection of individual spike-in tags is made according to the particular requirements of the combination of origin, preparation, and processing of the sample. Preferably, spike-in tags can be made by generating sequences to make up the majority of the tag comprising a sequence not known

25   to be present in any naturally occurring nucleic acid or to encode any naturally occurring protein. This is done to reduce the likelihood that the spike-in tag will be cross-reactive. Cross-reactivity indicates that a biopolymer has the ability to interact (*e.g.*, hybridize or bind) with more than one other biopolymer present during sample processing. For example, during sample processing using an oligonucleotide microarray, if the spike-in tag hybridizes

30   with its complementary probe as well as with a sequence in the biological sample or with a probe complementary to a sequence in the biological sample then the probe is said to be cross-reactive. Should a spike-in tag be highly cross-reactive, data collected from sample processing could be altered with respect to data obtained in the absence of the spike-in tag. In one embodiment, the sequences can initially be made by generating random sequences.

35   These sequences can then be assayed for their cross-reactivity with the biological sample to

be processed or probes designed to detect naturally occurring sequences in the biological sample during processing. Spike-in tags that are not cross-reactive with either the type of biopolymer being assayed in the sample or probes used to detect the biopolymers in the sample can be used for sample identification and tracking. This is done to ensure that the

5   spike-in tags will not affect the data received from sample processing and that they will solely serve as an internal identification of the sample. The random sequences are biopolymer residues (*e.g.*, nucleic acid or amino acid residues) that are generated without a preplanned specific design as to the actual resulting sequence, *i.e.*, when a monomer (*e.g.*, nucleotide, amino acid) is said to be random it is unpredictable what monomer will occur at

10 that residue. The random sequences can be synthesized by an unbiased synthesis scheme wherein each possible residue has an equal chance of being incorporated into the biopolymer at each position. Alternatively, the random sequences can be synthesized by a biased synthesis scheme wherein certain positions in the biopolymer have an increased chance of having one residue over another. Additionally, a combination of unbiased and

15 biased synthesis methods can be used to synthesize any one biopolymer. Predetermined sequences on either end or at internal positions may be added to the spike-in tag for the purposes of facilitating standard molecular biological manipulations. The predetermined sequences added at the ends of the spike-in tag are preferably less than 50%, less than 25%, less than 20%, less than 15%, less than 10%, or less than 5% of the total sequence of the

20 spike-in tag. Once generated, the spike-in tag sequence is determined and recorded for future comparison. Preferably, the spike-in tag comprises a sequence of at least 6-200 monomers.

The spike-in tags can be made of any type of macromolecule; preferably the molecular nature of a spike-in tag is consistent with that of the sample, that is, it is the same

25 type of molecule (*e.g.*, biopolymer) to be assayed in the sample as a result of the processing steps to which the sample will be subjected. For example, spike-in tags can be nucleic acids (*i.e.*, DNA or RNA), polypeptides, glycans, oligosaccharides, or small organic molecules.

In one embodiment the spike-in tag is a nucleic acid. For each nucleic acid spike-in tag, preferably there is available a complementary sequence such that contacting

30 the processed sample with a nucleic acid comprising the complementary sequence and detecting hybridization between the two, detects the presence and identity of the nucleic acid spike-in tag.

Nucleic acid spike-in tags of the invention are typically between about 20-5000 nucleic acid residues, more typically between about 40-80 nucleic acid residues, and

35 most typically between about 55-65 nucleic acid residues. In a preferred embodiment, using

inkjet arrays, the nucleic acid spike-in tag has 60 nucleic acid residues with 8 nucleic acid residues coming from predetermined sequences.

In another embodiment, the spike-in tag is a polypeptide. Each polypeptide spike-in tag preferably contains a binding moiety for which a binding partner is available

5   such that contacting the processed sample with said binding partner and detecting binding, detects the presence and identity of the polypeptide spike-in tag. In one embodiment, the binding moiety is an epitope recognized by an antibody, preferably a monoclonal antibody. Preferably, epitopes are unique (i.e., not endogenously expressed in cells or tissues that provide protein material for the samples) to minimize cross-reactivity of the antibodies

10   directed to spike-in tag epitopes with sample epitopes during detection.

Polypeptide spike-in tags of the invention are typically between 6-1500 amino acid residues, more typically between 10-50 amino acid residues, and most typically between 15-25 amino acid residues.

15   5.1.2 CONSTRUCTION AND ISOLATION

The process of preparing spike-in tags according to the invention can be performed in a number of ways. For example, spike-in tags can be prepared by standard synthetic techniques, e.g., using an automated synthesizer. An alternative method for preparing spike-in tags makes use of standard recombinant molecular biology techniques.
20

                               5.1.2.1 NUCLEIC ACIDS

In one embodiment the spike-in tag is a nucleic acid. Nucleic acid spike-in tags can be made up of deoxyribose nucleic acids, ribose nucleic acids, or analogues thereof; including, but not limited to, single-stranded cDNA (ss cDNA), double-stranded

25   cDNA (ds cDNA), mRNA, cRNA, and peptide nucleic acids (see e.g., Egholm et al., 1993, Nature 363:566-568; U.S. Patent No. 5,539,083). Nucleic acid spike-in tags of the invention are typically between about 20-5000 nucleic acid residues, more typically between about 40-80 nucleic acid residues, and most typically between about 55-65 nucleic acid residues. In a preferred embodiment, when the spike-in tags of the invention are ribonucleic

30   acids, the spike-in tags have a poly A tail, generally consisting of in the range of 20-250 contiguous adenine residues. In another specific embodiment, an RNA spike-in tag is capped at the 5' end with a 7-methylguanosine residue, and may optionally have a poly A tail.

In specific embodiments, spike-in tags of the invention can be made directly

35   through the use of chemical synthesis techniques, or they can be made with standard cloning

techniques (*e.g.*, as described in Sambrook et al., eds., *Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).

Plasmids in a host cell in which the plasmids can replicate can be used to
5   amplify a DNA sequence of choice that has been inserted into the plasmid. This technique can be used to obtain large quantities of DNA of interest from the amplified plasmid, for example by use of restriction enzymes and purification of the fragment of choice. DNA spike-in tags can be produced in this way (alternatively, the plasmid itself can be a spike-in tag, optionally after linearization). For example, cloning techniques can be utilized to create
10   plasmids containing a nucleic acid of the invention (or a portion thereof). In one embodiment, the plasmid comprising a nucleic acid of the invention is an expression plasmid. Expression plasmids are capable of directing the expression of genes to which they are operably linked. This means that the recombinant expression plasmids include one or more regulatory sequences, selected on the basis of the host cells to be used for
15   expression, which is operably linked to the nucleic acid sequence to be expressed. Within a recombinant expression plasmid, "operably linked" is intended to mean that the nucleotide sequence of interest is linked to the regulatory sequence(s) in a manner which allows for expression of the nucleotide sequence (*e.g.*, in an *in vitro* transcription/translation system or in a host cell when the plasmid is introduced into the host cell; a host cell being any
20   prokaryotic or eukaryotic cell capable of replicating the plasmid). The term "regulatory sequence" is intended to include promoters, enhancers and other expression control elements (*e.g.*, polyadenylation signals). Such regulatory sequences are described, for example, in Goeddel, *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, CA (1990). Regulatory sequences include those which direct
25   constitutive expression of a nucleotide sequence in many types of host cell and those which direct expression of the nucleotide sequence only in certain host cells (*e.g.*, tissue-specific regulatory sequences). It will be appreciated by those skilled in the art that the design of the expression plasmid can depend on such factors as the choice of the host cell to be transformed, the level of expression desired, etc. The expression plasmids comprising a
30   nucleic acid of the invention can be introduced into host cells to thereby produce RNA transcripts for use as spike-in tags.

Plasmid DNA can be introduced into prokaryotic or eukaryotic cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized
35   techniques for introducing foreign nucleic acid into a host cell, including calcium phosphate

or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, or electroporation. Suitable methods for transforming or transfecting host cells can be found in Sambrook, et al. (*supra*), and other laboratory manuals.

5   In a preferred embodiment, the recombinant expression plasmid can be transcribed *in vitro*, for example using T7 promoter regulatory sequences and T7 polymerase or using a SP6 promoter regulatory sequence and SP6 polymerase. Using this technique, RNA spike-in tags can be produced in isolation from other RNA molecules endogenously produced in the host cell.

       Standard techniques for isolation of nucleic acids can be used to isolate the
10   nucleic acid spike-in tags. In a preferred embodiment, the spike-in tags are RNA. Any RNA isolation technique which does not select against the isolation of mRNA can be utilized for the purification of such RNA samples (*see, e.g.*, Ausubel, F. M. et al., eds., 1987-1993, *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc. New York). Additionally, large numbers of cell samples can readily be processed using techniques well
15   known to those of skill in the art, such as, for example, the single-step RNA isolation process of Chomczynski (1989, U.S. Pat. No. 4,843,155).

5.1.2.2 POLYPEPTIDES

       In another embodiment, the spike-in tag is a polypeptide. Polypeptide spike-
20   in tags of the invention are typically between about 6-100 amino acid residues, more typically between about 10-50 amino acid residues, and most typically between about 15-25 amino acid residues. Polypeptide spike-in tags of the invention can be made directly through the use of chemical synthesis techniques, or they can be made with standard cloning techniques (*e.g.*, as described in Sambrook et al., eds., *Molecular Cloning: A Laboratory*
25   *Manual, 2nd ed., Cold Spring Harbor Laboratory*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).

       Standard cloning techniques utilize host cells containing expression plasmids that express the polypeptide of the invention. Such host cells and expression plasmids can be made according to methods described above. When polypeptide spike-in tags are
30   produced in this manner, the corresponding nucleic acid insert in the expression plasmid is typically between about 18-300 nucleic acid residues, more typically between about 30-150 nucleic acid residues, and most typically between about 45-75 nucleic acid residues.

       Standard techniques for isolation of polypeptides can be used to isolate the polypeptide spike-in tags.
35

5.2   ADDITION OF MOLECULAR BARCODES TO SAMPLES

The spike-in tags of the molecular barcode can be added at any time during the collection or processing of a sample. In one embodiment, the spike-in tags are added at the time of sample collection. In this way, the opportunities for misidentification or

5   handling errors are decreased.

A record is kept of the molecular barcode that corresponds to each sample for later sample identity verification. The record of molecular barcodes is a compendium of more than one and preferably more than 10, 50, 100, 200, 500, or 1000 molecular barcodes with each molecular barcode associated with sample identity. The compendium is

10   preferably stored on a computer.

This embodiment of the present invention lends itself to advance preparation of sample collection vessels, *i.e.*, tubes, bottles, slides, etc., containing pre-selected combinations of spike-in tags. For example, a sample collection tube may be coated on the inside surface with a number of spike-in tags. A label on the outside of the tube can provide

15   the nature of the molecular barcode already present. Upon addition of sample into the collection tube, the molecular barcode can instantly be recorded. This is particularly advantageous when collecting samples in a remote area or in the field (such as a corn or other crop field) away from a carefully controlled clinical setting. After processing and analysis, the particular profile of spike-in tags shown to be present in the sample can be

20   compared with the molecular barcode added at the time of collection to confirm identification of the sample. The molecular nature of the spike-in tag must be compatible with the type of processing to be performed

In another embodiment, the spike-in tags of the molecular barcode are added during sample processing. Because the spike-in tags are added during processing, the

25   molecular nature of the spike-in tags must be compatible with the type of processing to be performed in subsequent stages of the protocol yet to be performed. In one embodiment, sample preparation includes the synthesis of cRNA from total cellular RNA or total cellular mRNA. As schematically depicted in Figure 2A, sample processing may involve steps through different nucleic acid intermediates (*i.e.*, RNA, ss cDNA, ds cDNA, cRNA).

30   Nucleic acid spike-in tags can be added at any step in the process; preferably the spike-in tags are of the same molecular type that is predominant in the sample at the time of addition, or that is the molecule, or gives rise to the molecule, to be assayed as the output upon completion of sample processing. For example, spike-in tags added to a sample between steps 3 and 4 of the depicted protocol are preferably added as ds cDNA. In a

35   specific embodiment, when the sample processing includes the synthesis of cRNA from

total cellular RNA or total cellular mRNA, the spike-in tags are synthetic mRNA molecules and are added to the biological sample prior to the start of sample processing.

5.3 STRATEGIES OF SPIKE-IN TAG USE AS MOLECULAR BAR CODES

5   While any unique spike-in tag could, in principle, be used as a molecular barcode, the cost associated with synthesizing large numbers of unique spike-in tags (*i.e.*, one for each sample) is likely prohibitive. Thus, it is desirable to have a method of introducing a large number of unique molecular barcodes based upon a small collection of spike-in tags.

10   More than one individual spike-in tag may be added to a sample to create the molecular barcode. For example, the number of spike-in tags to be added to each biological sample is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more. In one embodiment, a sample may be identified by positional information when in a group of samples. For example, in this case, each row and column of a sample-containing plate is assigned its own unique spike-in tag.

15   With the addition of two spike-in tags to each sample (one corresponding to the row (or x-axis) and one corresponding to the column (or y-axis) of the sample-containing plate), a sample's molecular barcode can impart positional information. In this way, 96 samples can be identified with the creation of just 20 unique spike-in tags. Figure 6A illustrates this type of identification scheme. Additionally, molecular barcodes can be used to identify samples

20   by positional information in more than two dimensions. For example, three spike-in tags can be added to each sample, wherein each of the spike-in tags represents the sample's position on either the x-, y-, or z-axis.

In another embodiment, an alternate identification scheme for molecular barcodes is through the use of sample numbers (wherein each number uniquely identifies a

25   sample). A unique spike-in tag is assigned to the numbers 0-9 for each digit of a sample number. Thus ten different spike-in tag molecules are needed for each order of magnitude. In this way, 999 samples can be identified with the creation of just 30 unique spike-in tags. Figure 6B illustrates this type of identification scheme. The upper panel is an example of a key wherein each digit of the sample number is assigned a unique spike-in tag for the

30   numbers 0-9. The lower panel is an example of four different molecular barcodes that can be given to four different samples. In this example, three spike in tags are added to each sample to make the molecular barcode. In other embodiments, N spike-in tags are added to a sample wherein each of the spike-in tags represents a different digit in a N-digit identification number represented by the molecular barcode (wherein N is a whole number

35   that can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more).

The method of the present invention can be used to track a plurality of samples (*e.g.*, more than 1, 10, 50, 100, 200, 500, 1000, or 10000). Preferably, each sample originates from a different biological specimen. In one embodiment, each sample being tracked in the plurality is processed in the same manner.

5

5.4 MOLECULAR BARCODE USE IN SAMPLE TRACKING

Once a molecular barcode made up of spike-in tags has been added to a sample, the sample can be processed to obtain the desired data. In a preferred embodiment, the processing step comprises a plurality of serial transfers of said sample to different

10 containers or solid phases. In the present invention, the sample is preferably from isolated cell(s). In a more preferred embodiment, the isolated cells are isolated from a patient. A preferred embodiment for sample processing is on a microarray. The probes on microarrays can be any one of a number of different biopolymers, *e.g.*, DNAs, RNAs, peptide nucleic acids (PNAs) (see *e.g.*, Egholm *et al.*, 1993, *Nature 363*:566-568; U.S.

15 Patent No. 5,539,083), or polypeptides.

5.4.1 NUCLEOTIDE MICROARRAYS

In a preferred embodiment in the present invention, sample processing is through hybridization on a nucleotide microarray. In a more preferred embodiment, the

20 microarray is an oligonucleotide array. In a most preferred embodiment, the oligonucleotide array is used to monitor the expression levels of a large number of genetic transcripts at any one time. As a result, the spike-in tag in the most preferred embodiment is a mRNA with a poly A tail that is added to the sample prior to a processing step comprising conversion to cDNA.

25   Hybridization levels are preferably measured using polynucleotide probe arrays or microarrays. On a polynucleotide array, polynucleotide probes comprising sequences of interest are immobilized to the surface of a support, *e.g.*, a solid support. For example, the probes may comprise DNA sequences, RNA sequences, or copolymer sequences of DNA and RNA. The polynucleotide sequences of the probes may also

30 comprise DNA and/or RNA analogues (*e.g.*, peptide nucleic acids), or combinations thereof. For example, the polynucleotide sequences of the probe may be full or partial sequences of genomic DNA or mRNA derived from cells, or may be cDNA or cRNA sequences derived therefrom. The polynucleotide sequences of the probes may also be synthetic nucleotide sequences, such as synthetic oligonucleotide sequences. The probe

35

sequences can be synthesized either enzymatically *in vivo*, enzymatically *in vitro* (*e.g.*, by PCR), or non-enzymatically *in vitro*.

5   The probe or probes used in the methods of the invention are preferably immobilized to a solid support or surface which may be either porous or non-porous. For example, the probes of the invention may be polynucleotide sequences which are attached to a nitrocellulose or nylon membrane or filter. Such hybridization probes are well known in the art (see, *e.g.*, Sambrook et al., Eds., 1989, *Molecular Cloning: A Laboratory Manual, Vols. 1-3*, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York). Alternatively, the solid support or surface may be a glass or plastic surface.

10

5.4.1.1 HYBRIDIZATION ASSAY USING MICROARRAYS

        A microarray is an array of positionally-addressable binding (e.g., hybridization) sites on a support. Each of such binding sites comprises a plurality of polynucleotide molecules of a probe bound to the predetermined region on the support.

15 Microarrays can be made in a number of ways, of which several are described herein below. However produced, microarrays share certain characteristics. The arrays are preferably reproducible, allowing multiple copies of a given array to be produced and easily compared with each other. Preferably, the microarrays are made from materials that are stable under binding (*e.g.*, nucleic acid hybridization) conditions. The microarrays are preferably small,

20 *e.g.*, between about 1 cm^2 and 25 cm^2, preferably about 1 to 3 cm^2. However, both larger and smaller arrays are also contemplated and may be preferable, *e.g.*, for simultaneously evaluating a very large number of different probes.

        In a particularly preferred embodiment, hybridization levels are measured to microarrays of probes consisting of a solid phase on the surface of which are immobilized a

25 population of polynucleotides, such as a population of DNA or DNA mimics or, alternatively, a population of RNA or RNA mimics. The solid phase may be a nonporous or, optionally, a porous material such as a gel. Microarrays can be employed, *e.g.*, for analyzing the transcriptional state of a cell such as the transcriptional states of cells exposed to graded levels of a drug of interest or to graded perturbations to a biological pathway of

30 interest. Microarrays can be used to simultaneously screen a plurality of different probes to evaluate, *e.g.*, each probe's sensitivity and specificity for a particular target polynucleotide.

        Preferably, a given binding site or unique set of binding sites on the microarray will specifically bind (*e.g.*, hybridize) to the product of a single gene or gene transcript from a cell or organism (*e.g.*, to a specific mRNA or to a specific cDNA derived

35

therefrom). However, in general, other related or similar sequences may cross hybridize to a given binding site.

The microarrays used in the methods and compositions of the present invention include one or more test probes, each of which has a polynucleotide sequence that is complementary to a subsequence of RNA or DNA to be detected. Each probe preferably has a different nucleic acid sequence, and the position of each probe on the solid surface of the array is preferably known. Indeed, the microarrays are preferably addressable arrays, more preferably positionally addressable arrays. More specifically, each probe of the array is preferably located at a known, predetermined position on the solid support such that the identity (*i.e.*, the sequence) of each probe can be determined from its position on the array (*i.e.*, on the support or surface).

Preferably, the density of probes on a microarray is about 100 different (*i.e.*, non-identical) probes per 1 cm^2 or higher. More preferably, a microarray used in the methods of the invention will have at least 550 probes per 1 cm^2, at least 1000 probes per 1 cm^2, at least 1500 probes per 1 cm^2 or at least 2000 probes per 1 cm^2. In a particularly preferred embodiment, the microarray is a high density array, preferably having a density of at least about 2500 different probes per 1 cm^2. The microarrays used in the invention therefore preferably contain at least 2500, at least 5000, at least 10000, at least 15000, at least 20000, at least 25000, at least 50000 or at least 55000 different (*i.e.*, non-identical) probes. A subset of these probes will correspond to spike-in tags which may have ben added to the sample.

Such polynucleotides are preferably of the length of 15 to 200 bases, more preferably of the length of 20 to 100 bases, most preferably 40-60 bases. It will be understood that each probe sequence may also comprise linker sequences in addition to the sequence that is complementary to its target sequence. As used herein, a linker sequence refers to a sequence between the sequence that is complementary to its target sequence and the surface.

In one embodiment, the microarray is an array (*i.e.*, a matrix) in which each position represents a discrete binding site for a transcript encoded by a gene (*e.g.*, for an mRNA or a cDNA derived therefrom). For example, in various embodiments, the microarrays of the invention can comprise binding sites for products encoded by fewer than 50% of the genes in the genome of an organism. Alternatively, the microarrays of the invention can have binding sites for the products encoded by at least 50%, at least 75%, at least 85%, at least 90%, at least 95%, at least 99% or 100%, or at least 50, 100, 500, 1000, or 10000 of the genes in the genome of an organism. In other embodiments, the

microarrays of the invention can having binding sites for products encoded by fewer than 50%, by at least 50%, by at least 75%, by at least 85%, by at least 90%, by at least 95%, by at least 99% or by 100% of the genes expressed by a cell of an organism. The binding site can be a DNA or DNA analog to which a particular RNA can specifically hybridize. The

5   DNA or DNA analog can be, *e.g.*, a synthetic oligomer or a gene fragment, e.g. corresponding to an exon.

Preferably, the microarrays used in the invention have binding sites (*i.e.*, probes) for sets of genes for one or more genes relevant to the action of a drug of interest or in a biological pathway of interest. As discussed above, a "gene" is identified as a portion

10   of DNA that is transcribed by RNA polymerase, which may include a 5' untranslated region (UTR), introns, exons and a 3' UTR. The number of genes in a genome can be estimated from the number of mRNA molecules expressed by the cell or organism, or by extrapolation of a well characterized portion of the genome. When the genome of the organism of interest has been sequenced, the number of open reading frames (ORFs) can be determined and

15   mRNA coding regions identified by analysis of the DNA sequence. For example, the genome of *Saccharomyces cerevisiae* has been completely sequenced and is reported to have approximately 6275 ORFs encoding sequences longer the 99 amino acid residues in length. Analysis of these ORFs indicates that there are 5,885 ORFs that are likely to encode protein products (Goffeau et al., 1996, *Science 274*:546-567). In contrast, the

20   human genome is estimated to contain approximately 30000 to 130000 genes (see Crollius et al., 2000, *Nature Genetics* 25:235-238; Ewing et al., 2000, *Nature Genetics* 25:232-234). Genome sequences for other organisms, including but not limited to *Drosophila, C. elegans*, plants, *e.g.*, rice and Arabidopsis, and mammals, *e.g.*, mouse and human, are also completed or nearly completed. Thus, in preferred embodiments of the invention, array set

25   comprising probes for all genes in the genome of an organism is provided.

It will be appreciated that when a sample of target nucleic acid molecules, *e.g.*, cDNA complementary to the RNA of a cell is made and hybridized to a microarray under suitable hybridization conditions, the level of hybridization to the site in the array will reflect the prevalence of the corresponding complementary sequences in the sample. For

30   example, when detectably labeled (*e.g.*, with a fluorophore) cDNA is hybridized to a microarray, the site on the array corresponding to a nucleotide sequence that is not in the sample will have little or no signal (*e.g.*, fluorescent signal), and a nucleotide sequence that is prevalent in the sample will have a relatively strong signal. The relative abundance of different nucleotide sequences in a sample may be determined by the signal strength pattern

35   of probes on a microarray. Because a known quantity of spike-in tags can be added to a

sample, the resulting level of molecular barcode signal after hybridization can be used as a control to allow comparison across multiple microarray scans.

Nucleic acids from samples from two different cells subjected to two different conditions can be hybridized to the binding sites of the microarray using a two-

5   color protocol. In the case of drug responses, one cell sample is exposed to a drug and another cell sample of the same type is not exposed to the drug. The cDNA derived from each of the two cell types is differently labeled (*e.g.*, with Cy3 and Cy5) so that they can be distinguished. In one embodiment, for example, cDNA from a cell treated with a drug (or having a mutation or a disease, etc.) is synthesized using a fluorescein-labeled dNTP, and

10   cDNA from a second cell, not drug-exposed, is synthesized using a rhodamine-labeled dNTP. When the two cDNA molecules are mixed and hybridized to the microarray, the relative intensity of signal from each cDNA set is determined for each site on the array, and any relative difference in abundance of a particular gene detected.

In the example described above, the nucleic acid from the drug-treated cell

15   will fluoresce green when the fluorophore is stimulated and the nucleic acid from the untreated cell will fluoresce red. As a result, when the drug treatment has no effect, either directly or indirectly, on the transcription of a particular gene in a cell, the expression patterns will be indistinguishable in both cells and, upon reverse transcription, red-labeled and green-labeled nucleic acids will be equally prevalent. When hybridized to the

20   microarray, the binding site(s) for that species of nucleic acid will emit wavelengths characteristic of both fluorophores. In contrast, when the drug-exposed cell is treated with a drug that, directly or indirectly, change the transcription of a particular gene in the cell, the expression pattern as represented by ratio of green to red fluorescence for each binding site will change. When the drug increases the prevalence of an mRNA, the ratios for each

25   binding site of the mRNA will increase, whereas when the drug decreases the prevalence of an mRNA, the ratio for each for each binding site in the mRNA will decrease.

The use of a two-color fluorescence labeling and detection scheme to define alterations in gene expression has been described in connection with detection of mRNA molecules, *e.g.*, in Shena et al., 1995, Quantitative monitoring of gene expression patterns

30   with a complementary DNA microarray, *Science* 270:467-470. An advantage of using cDNA labeled with two different fluorophores is that a direct and internally controlled comparison of the mRNA or exon expression levels corresponding to each arrayed gene in two cell states can be made, and variations due to minor differences in experimental conditions (*e.g.*, hybridization conditions) will not affect subsequent analyses. However, it

35   will be recognized that it is also possible to use cDNA from a single cell, and compare, for

example, the absolute amount of a particular exon in, *e.g.*, a drug-treated or pathway-perturbed cell and an untreated cell. Furthermore, labeling with more than two colors is also contemplated in the present invention. In some embodiments of the invention, at least 5, 10, 20, or 100 dyes of different colors can be used for labeling. Such labeling

5   permits simultaneous hybridizing of the distinguishably labeled cDNA populations to the same array, and thus measuring, and optionally comparing the expression levels of, mRNA molecules derived from more than two samples. Dyes that can be used include, but are not limited to, fluorescein and its derivatives, rhodamine and its derivatives, texas red, 5'carboxy-fluorescein (FMA), 2',7'-dimethoxy-4',5'-dichloro-6-carboxy-fluorescein (JOE),

10 N,N,N',N'-tetramethyl-6-carboxy-rhodamine (TAMRA), 6'carboxy-X-rhodamine (ROX), HEX, TET, IRD40, and IRD41, cyanine dyes, including but are not limited to Cy3, Cy3.5 and Cy5; BODIPY dyes including but are not limited to BODIPY-FL, BODIPY-TR, BODIPY-TMR, BODIPY-630/650, and BODIPY-650/670; and ALEXA dyes, including but are not limited to ALEXA-488, ALEXA-532, ALEXA-546, ALEXA-568, and ALEXA-

15 594; as well as other fluorescent dyes which will be known to those who are skilled in the art.

5.4.1.2 PREPARING PROBES FOR MICROARRAYS

As noted above, the probe to which a particular polynucleotide molecule

20 specifically hybridizes is a complementary polynucleotide sequence. The probes may comprise DNA corresponding to a portion of each gene in an organism's genome. In one embodiment, the probes of the microarray are complementary RNA. DNA can be obtained, *e.g.*, by polymerase chain reaction (PCR) amplification of exon segments from genomic DNA, cDNA (*e.g.*, by RT-PCR), or cloned sequences. PCR primers are preferably chosen

25 based on known sequence of the exons or cDNA that result in amplification of unique fragments (*i.e.*, fragments that do not share more than 10 bases of contiguous identical sequence with any other fragment on the microarray). Computer programs that are well known in the art are useful in the design of primers with the desired specificity and optimal amplification properties, such as *Oligo* version 5.0 (National Biosciences). Typically each

30 probe on the microarray will be between 20 bases and 600 bases, and usually between 30 and 200 bases in length. PCR methods are well known in the art, and are described, for example, in Innis et al., eds., 1990, *PCR Protocols: A Guide to Methods and Applications*, Academic Press Inc., San Diego, CA. It will be apparent to one skilled in the art that controlled robotic systems are useful for isolating and amplifying nucleic acids.

35

In alternative embodiments, the hybridization sites (*i.e.*, the probes) are made from plasmid or phage clones of genes, cDNA molecules (*e.g.*, expressed sequence tags), or inserts therefrom (Nguyen et al., 1995, *Genomics 29*:207-209).

In the preferred embodiment, the means for generating the polynucleotide
5   probes of the microarray is by synthesis of synthetic polynucleotides or oligonucleotides, *e.g.*, using N-phosphonate or phosphoramidite chemistries (Froehler et al., 1986, *Nucleic Acid Res. 14*:5399-5407; McBride et al., 1983, *Tetrahedron Lett. 24*:246-248). Synthetic sequences are typically between about 15 and about 600 bases in length, more typically between about 20 and about 100 bases, most preferably between about 40 and about 70
10   bases in length.

The probes on the microarrays are macromolecules attached to the solid support of a microarray. In the present invention, the probes are preferably nucleic acid sequences (or fragments thereof). Included in the probes on the microarray preferably are probes capable of hybridizing to each of the possible the spike-in tags. In one embodiment,
15   the probes hybridizing to the spike-in tags in the processed sample are the compliment of the nucleic acid sequence of each possible spike-in tag. In another embodiment, the probes hybridizing to the spike-in tags in the processed sample are the exact nucleic acid sequence of each possible spike-in tag. This is to ensure that the particular profile of spike-in tags that make up the molecular barcode of a sample will be detectable during microarray
20   processing.

5.4.1.3 ATTACHING PROBES TO THE SOLID SURFACE

Preformed polynucleotide probes can be deposited on a support to form the array. Alternatively, polynucleotide probes can be synthesized directly on the support to
25   form the array. The probes are attached to a solid support or surface, which may be made, *e.g.*, from glass, plastic (*e.g.*, polypropylene, nylon), polyacrylamide, nitrocellulose, gel, or other porous or nonporous material.

A preferred method for attaching the nucleic acids to a surface is by printing on glass plates, as is described generally by Schena *et al*, 1995, *Science 270*:467-470. This
30   method is especially useful for preparing microarrays of cDNA (See also, DeRisi *et al*, 1996, *Nature Genetics 14*:457-460; Shalon et al., 1996, *Genome Res. 6*:639-645; and Schena et al., 1995, *Proc. Natl. Acad. Sci. U.S.A. 93*:10539-11286).

A second preferred method for making microarrays is by making high-density oligonucleotide arrays. Techniques are known for producing arrays containing
35   thousands of oligonucleotides complementary to defined sequences, at defined locations on

a surface using photolithographic techniques for synthesis *in situ* (see, Fodor et al., 1991, *Science 251*:767-773; Pease et al., 1994, *Proc. Natl. Acad. Sci. U.S.A. 91*:5022-5026; Lockhart et al., 1996, *Nature Biotechnology 14*:1675; U.S. Patent Nos. 5,578,832; 5,556,752; and 5,510,270) or other methods for rapid synthesis and deposition of defined

5   oligonucleotides (Blanchard et al., *Biosensors & Bioelectronics 11*:687-690). When these methods are used, oligonucleotides (*e.g.*, 60-mers) of known sequence are synthesized directly on a surface such as a derivatized glass slide. The array produced can be redundant, with several oligonucleotide molecules per gene.

Other methods for making microarrays, *e.g.*, by masking (Maskos and

10 Southern, 1992, *Nucl. Acids. Res. 20*:1679-1684), may also be used. In principle, and as noted *supra*, any type of array, for example, dot blots on a nylon hybridization membrane (see Sambrook et al., *supra*) could be used. However, as will be recognized by those skilled in the art, very small arrays will frequently be preferred because hybridization volumes will be smaller.

15   In a particularly preferred embodiment, microarrays of the invention are manufactured by means of an ink jet printing device for oligonucleotide synthesis, *e.g.*, using the methods and systems described by Blanchard in International Patent Publication No. WO 98/41531, published September 24, 1998; Blanchard et al., 1996, *Biosensors and Bioelectronics 11*:687-690; Blanchard, 1998, in *Synthetic DNA Arrays in Genetic*

20 *Engineering*, Vol. 20, J.K. Setlow, Ed., Plenum Press, New York at pages 111-123; Hughes et al., 2001, *Nature Biotechnology* 19:342-347; and U.S. Patent No. 6,028,189 to Blanchard. Specifically, the oligonucleotide probes in such microarrays are preferably synthesized in arrays, *e.g.*, on a glass slide, by serially depositing individual nucleotide bases in microdroplets of a high surface tension solvent such as propylene carbonate. The

25 microdroplets have small volumes (*e.g.*, 100 pL or less, more preferably 50 pL or less) and are separated from each other on the microarray (*e.g.*, by hydrophobic domains) to form circular surface tension wells which define the locations of the array elements (*i.e.*, the different probes). Polynucleotide probes are attached to the surface covalently at the 3' end of the polynucleotide.

30

5.4.1.4 TARGET POLYNUCLEOTIDE MOLECULES

Target polynucleotides are the polynucleotides of the biological samples that are being tracked through the processing steps that occur during analysis of the biological sample. Target polynucleotides can be RNA molecules such as, but by no means limited to

35 messenger RNA (mRNA) molecules, ribosomal RNA (rRNA) molecules, cRNA molecules

(*i.e.*, RNA molecules prepared from cDNA molecules that are transcribed *in vitro*) and fragments thereof. Additionally, target polynucleotides may also be, but are not limited to, DNA molecules such as genomic DNA molecules, cDNA molecules, and fragments thereof including oligonucleotides, ESTs, STSs, *etc.* In specific embodiments, the sample

5   comprises more than 1000, 5000, 10000, 50000, or 100000 nucleic acid molecules of different nucleotide sequences.

The target polynucleotides may be from any source. For example, the target polynucleotide molecules may be naturally occurring nucleic acid molecules such as genomic or extragenomic DNA molecules isolated from an organism, or RNA molecules,

10   such as mRNA molecules, isolated from an organism. Alternatively, the polynucleotide molecules may be synthesized, including, *e.g.*, nucleic acid molecules synthesized enzymatically *in vivo* or *in vitro*, such as cDNA molecules, or polynucleotide molecules synthesized by PCR, RNA molecules synthesized by *in vitro* transcription, *etc.* The sample of target polynucleotides can comprise, *e.g.*, molecules of DNA, RNA, or copolymers of

15   DNA and RNA. In preferred embodiments, the target polynucleotides of the invention will correspond to particular genes or to particular gene transcripts (*e.g.*, to particular mRNA sequences expressed in cells or to particular cDNA sequences derived from such mRNA sequences). However, in many embodiments, particularly those embodiments wherein the polynucleotide molecules are derived from mammalian cells, the target polynucleotides may

20   correspond to particular fragments of a gene transcript. For example, the target polynucleotides may correspond to different exons of the same gene, *e.g.*, so that different splice variants of that gene may be detected and/or analyzed.

In preferred embodiments, the target polynucleotides to be analyzed are prepared *in vitro* from nucleic acids extracted from cells. For example, in one embodiment,

25   RNA is extracted from cells (*e.g.*, total cellular RNA, poly(A)$^+$ messenger RNA, fraction thereof) and messenger RNA is purified from the total extracted RNA. Methods for preparing total and poly(A)$^+$ RNA are well known in the art, and are described generally, *e.g.*, in Sambrook et al., *supra.* In one embodiment, RNA is extracted from cells of the various types of interest in this invention using guanidinium thiocyanate lysis followed by

30   CsCl centrifugation and an oligo dT purification (Chirgwin et al., 1979, *Biochemistry 18*:5294-5299). In another embodiment, total RNA is extracted from cells using guanidinium thiocyanate lysis followed by purification on RNeasy columns (Qiagen). cDNA is then synthesized from the purified mRNA using, *e.g.*, oligo-dT or random primers. In preferred embodiments, the target polynucleotides are cRNA prepared from cDNA

35   prepared from purified mRNA or from total RNA extracted from cells. As used herein,

cRNA can either be complementary to (anti-sense) or of the same sequence (sense) as the sample RNA. The extracted RNA molecules are amplified using a process in which double-stranded cDNA molecules are synthesized from the sample RNA molecules using primers linked to an RNA polymerase promoter. As a result, RNA polymerase promoters

5   can be incorporated into either or both strands of the cDNA. Using the RNA polymerase promoter that is on the first strand of the cDNA molecule, cRNA can be synthesized that is the same sequence as the sample RNA. To synthesize cRNA complementary to the sample RNA, transcription can be initiated from the RNA polymerase promoter that is on the second strand of the double-stranded cDNA molecule using an RNA polymerase (see, e.g.,

10   U.S. Patent Nos. 5,891,636, 5,716,785; 5,545,522 and 6,132,997; see also, U.S. Patent No. 6,271,002 and U.S. Provisional Patent Application Serial No. 60/253,641, filed on November 28, 2000, by Ziman et al.). Both oligo-dT primers (U.S. Patent Nos. 5,545,522 and 6,132,997) or random primers (U.S. Provisional Patent Application Serial No. 60/253,641, filed on November 28, 2000, by Ziman et al.) that contain an RNA polymerase

15   promoter or complement thereof can be used. Preferably, the target polynucleotides are short and/or fragmented polynucleotide molecules which are representative of the original nucleic acid population of the cell. In one embodiment, total RNA is used as input for cRNA synthesis. An oligo-dT primer containing a T7 RNA polymerase promoter sequence can be used to prime first strand cDNA synthesis. When second strand synthesis is desired,

20   random hexamers can be used to prime second strand cDNA synthesis by a reverse transcriptase. This reaction yields a double-stranded cDNA that contains the T7 RNA polymerase promoter at the 3' end. The double-stranded cDNA can then be transcribed into cRNA by T7 RNA polymerase.

The target polynucleotides to be analyzed are preferably detectably labeled.

25   For example, cDNA can be labeled directly, e.g., with nucleotide analogs, or indirectly, e.g., by making a second, labeled cDNA strand using the first strand as a template. Alternatively, the double-stranded cDNA can be transcribed into cRNA and labeled.

Preferably, the detectable label is a fluorescent label, e.g., by incorporation of nucleotide analogs. Other labels suitable for use in the present invention include, but are

30   not limited to, biotin, imminobiotin, antigens, cofactors, dinitrophenol, lipoic acid, olefinic compounds, detectable polypeptides, electron rich molecules, enzymes capable of generating a detectable signal by action upon a substrate, and radioactive isotopes. Preferred radioactive isotopes include ^{32}P, ^{35}S, ^{14}C, ^{15}N and ^{125}I. Fluorescent molecules suitable for the present invention include, but are not limited to, fluorescein and its

35   derivatives, rhodamine and its derivatives, texas red, 5'carboxy-fluorescein (FMA), 2',7'-

dimethoxy-4',5'-dichloro-6-carboxy-fluorescein (JOE), N,N,N',N'-tetramethyl-6-carboxy-
rhodamine (TAMRA), 6'carboxy-X-rhodamine (ROX), HEX, TET, IRD40, and IRD41.
Fluorescent molecules that are suitable for the invention further include: cyamine dyes,
including by not limited to Cy3, Cy3.5 and Cy5; BODIPY dyes including but not limited to

5   BODIPY-FL, BODIPY-TR, BODIPY-TMR, BODIPY-630/650, and BODIPY-650/670;
and ALEXA dyes, including but not limited to ALEXA-488, ALEXA-532, ALEXA-546,
ALEXA-568, and ALEXA-594; as well as other fluorescent dyes which will be known to
those who are skilled in the art. Electron rich indicator molecules suitable for the present
invention include, but are not limited to, ferritin, hemocyanin, and colloidal gold.

10   Alternatively, in less preferred embodiments the target polynucleotides may be labeled by
specifically complexing a first group to the polynucleotide. A second group, covalently
linked to an indicator molecules and which has an affinity for the first group, can be used to
indirectly detect the target polynucleotide. In such an embodiment, compounds suitable for
use as a first group include, but are not limited to, biotin and imminobiotin. Compounds

15   suitable for use as a second group include, but are not limited to, avidin and streptavidin.

5.4.1.5 HYBRIDIZATION TO MICROARRAYS

As described *supra*, nucleic acid hybridization and wash conditions are
chosen so that the polynucleotide molecules to be analyzed (or target polynucleotide

20   molecules) specifically bind or specifically hybridize to the complementary polynucleotide
sequences of the array, preferably to one or more specific array sites, wherein its
complementary sequence is located.

Arrays containing double-stranded probe DNA situated thereon are
preferably subjected to denaturing conditions to render the DNA single-stranded prior to

25   contacting with the target polynucleotide molecules. Arrays containing single-stranded
probe DNA (*e.g.*, synthetic oligodeoxyribonucleic acids) may need to be denatured prior to
contacting with the target polynucleotide molecules, *e.g.*, to remove hairpins or dimers
which form due to self complementary sequences.

Optimal hybridization conditions will depend on the length (*e.g.*, oligomer

30   versus polynucleotide greater than 200 bases) and type (*e.g.*, RNA, or DNA) of probe and
target nucleic acids. General parameters for specific (*i.e.*, stringent) hybridization
conditions for nucleic acids are described in Sambrook et al., (*supra*), and in Ausubel et al.,
1987, *Current Protocols in Molecular Biology*, Greene Publishing and Wiley-Interscience,
New York. When the cDNA microarrays of Schena et al. are used, typical hybridization

35   conditions are hybridization in 5 X SSC plus 0.2% SDS at 65 °C for four hours, followed

by washes at 25 °C in low stringency wash buffer (1 X SSC plus 0.2% SDS), followed by 10 minutes at 25 °C in higher stringency wash buffer (0.1 X SSC plus 0.2% SDS) (Shena et al., 1996, *Proc. Natl. Acad. Sci. U.S.A. 93*:10614). Useful hybridization conditions are also provided in, *e.g.*, Tijessen, 1993, *Hybridization With Nucleic Acid Probes*, Elsevier Science

5   Publishers B.V. and Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press, San Diego, CA.

Particularly preferred hybridization conditions for use with the screening and/or signaling chips of the present invention include hybridization at a temperature at or near the mean melting temperature of the probes (*e.g.*, within 5 °C, more preferably within

10   2 °C) in 1 M NaCl, 50 mM MES buffer (pH 6.5), 0.5% sodium Sarcosine and 30% formamide.

5.4.1.6 SIGNAL DETECTION AND DATA ANALYSIS

It will be appreciated that when target sequences, *e.g.*, cDNA or cRNA,

15   complementary to the RNA of a cell is made and hybridized to a microarray under suitable hybridization conditions, the level of hybridization to the site in the array corresponding to a particular gene will reflect the prevalence in the cell of mRNA or mRNA molecules containing the transcript from that gene. For example, when detectably labeled (*e.g.*, with a fluorophore) cDNA complementary to the total cellular mRNA is hybridized to a

20   microarray, the site on the array corresponding to a gene (*i.e.*, capable of specifically binding the product or products of the gene expressing) that is not transcribed in the cell will have little or no signal (*e.g.*, fluorescent signal), and a gene for which the encoded mRNA expressing the transcript is prevalent will have a relatively strong signal.

In preferred embodiments, target sequences, *e.g.*, cDNA molecules or cRNA

25   molecules, from two different cells are hybridized to the binding sites of the microarray. In the case of drug responses one cell sample is exposed to a drug and another cell sample of the same type is not exposed to the drug. In the case of pathway responses one cell is exposed to a pathway perturbation and another cell of the same type is not exposed to the pathway perturbation. The cDNA or cRNA derived from each of the two cell types are

30   differently labeled so that they can be distinguished. In one embodiment, for example, cDNA from a cell treated with a drug (or otherwise perturbed) is synthesized using a fluorescein-labeled dNTP, and cDNA from a second cell, not drug-exposed, is synthesized using a rhodamine-labeled dNTP. When the two cDNA molecules are mixed and hybridized to the microarray, the relative intensity of signal from each cDNA set is

35

determined for each site on the array, and any relative difference in abundance of a particular transcript detected.

In the example described above in the previous paragraph, the cDNA from the drug-treated (or otherwise perturbed) cell will fluoresce green when the fluorophore is

5   stimulated and the cDNA from the untreated cell will fluoresce red. As a result, when the drug treatment has no effect, either directly or indirectly, on the transcription of a particular gene in a cell, the expression pattern will be indistinguishable in both cells and, upon reverse transcription, red-labeled and green-labeled cDNA will be equally prevalent. When hybridized to the microarray, the binding site(s) for that species of RNA will emit

10   wavelengths characteristic of both fluorophores. In contrast, when the drug-exposed cell is treated with a drug that, directly or indirectly, changes the transcription splicing of a particular gene in the cell, the expression pattern as represented by ratio of green to red fluorescence for each transcript binding site will change. When the drug increases the prevalence of an mRNA, the ratios for each transcript fragment expressed in the mRNA will

15   increase, whereas when the drug decreases the prevalence of an mRNA, the ratio for each exons expressed in the mRNA will decrease.

The use of a two-color fluorescence labeling and detection scheme to define alterations in gene expression has been described in connection with detection of mRNA molecules, e.g., in Shena et al., 1995, Quantitative monitoring of gene expression patterns

20   with a complementary DNA microarray, Science 270:467-470. An advantage of using target sequences, e.g., cDNA molecules or cRNA molecules, labeled with two different fluorophores is that a direct and internally controlled comparison of the mRNA expression levels corresponding to each arrayed gene in two cell states can be made, and variations due to minor differences in experimental conditions (e.g., hybridization conditions) will not

25   affect subsequent analyses. However, it will be recognized that it is also possible to use cDNA from a single cell, and compare, for example, the absolute amount of a particular exon in, e.g., a drug-treated or otherwise perturbed cell and an untreated cell.

In other preferred embodiments, single channel detection methods, e.g., using one-color fluorescence labeling, are used (see U.S. Patent Application Serial No.

30   09/781,814, filed on February 12, 2001). In this embodiment, arrays comprising reverse-complement (RC) probes are designed and produced. Because a reverse complement of a DNA sequence has sequence complexity that is equivalent to the corresponding forward-strand (FS) probe that is complementary to a target sequence with respect to a variety of measures (e.g., measures such as GC content and GC trend are invariant under the reverse

35   complement), a RC probe is used to as a control probe for determination of level of non-

specific cross hybridization to the corresponding FS probe. The significance of the FS probe intensity of a target sequence is determined by comparing the raw intensity measurement for the FS probe and the corresponding raw intensity measurement for the RC probe in conjunction with the respective measurement errors. In a preferred embodiment, a

5 transcript is called present if the intensity difference between the FS probe and the corresponding RC probe is significant. More preferably, a transcript is called present if the FS probe intensity is also significantly above background level. Single channel detection methods can be used in conjunction with multi-color labeling. In one embodiment, a plurality of different samples, each labeled with a different color, is hybridized to an array.

10 Differences between FS and RC probes for each color are used to determine the level of hybridization of the corresponding sample.

When fluorescently labeled probes are used, the fluorescence emissions at each site of a transcript array can be, preferably, detected by scanning confocal laser microscopy. In one embodiment, a separate scan, using the appropriate excitation line, is

15 carried out for each of the two fluorophores used. Alternatively, a laser can be used that allows simultaneous specimen illumination at wavelengths specific to the two fluorophores and emissions from the two fluorophores can be analyzed simultaneously (see Shalon et al., 1996, *Genome Res. 6*:639-645). In a preferred embodiment, the arrays are scanned with a laser fluorescence scanner with a computer controlled X-Y stage and a microscope

20 objective. Sequential excitation of the two fluorophores is achieved with a multi-line, mixed gas laser, and the emitted light is split by wavelength and detected with two photomultiplier tubes. Such fluorescence laser scanning devices are described, *e.g.*, in Schena et al., 1996, *Genome Res. 6*:639-645. Alternatively, the fiber-optic bundle described by Ferguson et al., 1996, *Nature Biotechnology 14*:1681-1684, may be used to monitor

25 mRNA abundance levels at a large number of sites simultaneously.

Signals are recorded and, in a preferred embodiment, analyzed by computer, *e.g.*, using a 12 bit or 16 bit analog to digital board. In one embodiment, the scanned image is despeckled using a graphics program (*e.g.*, Hijaak Graphics Suite) and then analyzed using an image gridding program that creates a spreadsheet of the average hybridization at

30 each wavelength at each site. If necessary, an experimentally determined correction for cross talk (or overlap) between the channels for the two fluors may be made. For any particular hybridization site on the transcript array, a ratio of the emission of the two fluorophores can be calculated. The ratio is independent of the absolute expression level of the cognate gene, but is useful for genes whose expression is significantly modulated by

35 drug administration, gene deletion, or any other tested event.

The relative abundance of an mRNA in two cells or cell lines is preferably scored as perturbed (*i.e.*, the abundance is different in the two sources of mRNA tested) or as not perturbed (*i.e.*, the relative abundance is the same). As used herein, a difference between the two sources of RNA of at least a factor of about 25% (*i.e.*, RNA is 25% more

5   abundant in one source than in the other source), more usually about 50%, even more often by a factor of about 2 (*i.e.*, twice as abundant), 3 (three times as abundant), or 5 (five times as abundant) is preferably scored as a perturbation.

It is, however, also advantageous to determine the magnitude of the relative difference in abundances for an mRNA expressed in an mRNA in two cells or in two cell

10   lines. This can be carried out, as noted above, by calculating the ratio of the emission of the two fluorophores used for differential labeling, or by analogous methods that will be readily apparent to those of skill in the art.

5.4.2 PROTEIN MICROARRAYS

15   In an embodiment in the present invention, sample processing is through binding on a protein microarray. In a preferred embodiment, the protein microarray is used to monitor the expression levels or activity levels of a large number of proteins at any one time. As a result, the spike-in tag in this embodiment is a polypeptide that is added to the sample prior to processing. Each polypeptide spike-in tag preferably has a corresponding

20   binding partner available on the protein array such that contacting the processed sample with said binding partner and detecting binding, detects the presence and identity of the polypeptide spike-in tag.

On a protein microarray, polypeptide probes possessing the ability to bind polypeptides of interest are immobilized to the surface of a support, *e.g.*, a solid support.

25   For example, polypeptide probes may be prepared using standard solid-phase techniques for the synthesis of peptides. As is generally known, polypeptides can be prepared using commercially available equipment and reagents following the manufacturers' instructions for blocking interfering groups, protecting the amino acid to be reacted, coupling, deprotection, and capping of unreacted residues. Alternatively, the polypeptide probes can

30   be made through the use of recombinant DNA technology (see methods described in Section 5.1.2). The polypeptide probes may contain non-peptide linkages and/or modified or non-naturally occurring amino acids, *e.g.*, D-amino acids, phosphorous analogs of amino acids, such as α-amino phosphoric acids and α-amino phosphoric acids.

The probe or probes used in the methods of the invention are preferably

35   immobilized to a solid support or surface which may be either porous or non-porous. For

example, the probes of the invention may be polypeptide sequences which are attached to a nitrocellulose or nylon membrane or filter. Alternatively, the solid support or surface may be a glass or plastic surface.

5   Purified polypeptides can be placed on a positionally addressable array with a plurality of polypeptides attached to a surface of a solid support, with each polypeptide being at a different position on the solid support, wherein the plurality of polypeptides comprises at least 10, 50, 100, 250, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 50000, or 100000 different polypeptides expressed in a single biological sample. A variety of attachment methods are known in the art. In one

10   embodiment, the proteins are printed onto the solid support (*e.g.*, Haab et al., 2001, *Genome Biology* 2:RESEARCH0004.1-RESEARCH0004.13).

In one embodiment, the probes on the protein microarrays capable of binding to each of the possible polypeptide spike-in tags in the sample are antibodies or fragments thereof attached to the solid support of a microarray. In more preferred embodiment, the

15   antibodies are monoclonal antibodies or fragments (*e.g.*, Fab fragments) thereof. This is to ensure that the particular profile of spike-in tags that make up the molecular barcode of a sample will be detectable during protein microarray processing.

It will be appreciated that when a sample of proteins is bound to a protein microarray under suitable conditions, the level of binding to a particular site in the array will

20   reflect the prevalence of the corresponding binding partner in the sample. The level of binding between polypeptide probe on the microarray and protein in the sample is preferably indicated by signaling compounds. For example, when a protein sample is bound to a protein microarray, the site on the array corresponding to a polypeptide probe with a corresponding binding partner not in the sample will have little or no signal, and a

25   polypeptide probe with a corresponding binding partner that is prevalent in the sample will have a relatively strong signal. The relative abundance of different proteins in a sample may be determined by the signal strength pattern of probes on a microarray. Because a known quantity of spike-in tags can be added to a sample, the resulting level of molecular barcode signal after hybridization can be used as a control to allow comparison across

30   multiple microarray scans. In one embodiment, one or more signal compounds (*e.g.*, fluorescent dyes) are directly attached to the proteins of the sample as well as the spike-in tags. In another embodiment, one or more signal compounds are attached to the proteins of the sample as well as the spike-in tags indirectly (*e.g.*, through the use of a fluorescently labeled antibodies).

35

The polypeptide probes on the protein microarray capable of binding to biopolymers in the sample (but not capable of binding to polypeptide spike-in tags) may or may not be antibodies or fragments thereof. In a specific embodiment, all of the probes on the microarray are antibodies or fragments or derivatives containing the binding domain

5   thereof.

5.4.3 ENZYMATIC MICROARRAYS

In an embodiment in the present invention, sample processing involves measurement of activity levels in an enzymatic microarray. In a preferred embodiment, the

10   enzymatic microarray is used to monitor the levels of a large number of enzymatic activities at any one time. In such embodiment, the spike-in tag is an enzyme substrate that is added to the sample prior to processing. In specific embodiments, the spike-in tag is a protein, an RNA, and/or a synthetic molecule comprising a sequence not known to be present in any naturally occurring nucleic acid or to encode any naturally occurring protein.

15   On an enzymatic microarray, enzyme probes with activities of interest are immobilized to the surface of a support, *e.g.*, a solid support. In one embodiment, enzymes that are protein may be prepared using standard solid-phase techniques for the synthesis of peptides. As is generally known, polypeptides can be prepared using commercially available equipment and reagents following the manufacturers' instructions for blocking

20   interfering groups, protecting the amino acid to be reacted, coupling, deprotection, and capping of unreacted residues. In another embodiment, enzymes that are nucleic acids (*e.g.*, ribozymes) can be made directly through the use of chemical synthesis techniques. Alternatively, either protein or nucleic acid enzyme probes can be made through the use of recombinant DNA technology (see methods described in Section 5.1.2).

25   The probe or probes used in the methods of the invention are preferably immobilized to a solid support or surface which may be either porous or non-porous. For example, the probes of the invention may be attached to a nitrocellulose or nylon membrane or filter. Alternatively, the solid support or surface may be a glass or plastic surface. The solid support may have cavities etched into it to provide a well for each enzymatic probe

30   (see *e.g.*, Curey et al., 2001, *Analytical Biochemistry* 293:178-184).

In one embodiment, one or more products of the enzymatic reaction can be detected directly. In another embodiment, one or more products of the enzymatic reaction interacts with a non-detectable molecule to alter it so that it is subsequently detectable (*e.g.*, nonfluorescent resazurin altered to fluorescent resorufin by hydride transfer).

35

In another embodiment, the enzymatic array can also include non-enzyme protein probes (see *e.g.*, Arenkov et al., 2000, *Analytical Biochemistry* 278:123-131). Probes capable of interacting with the spike-in tags can be either enzymes or non-enzyme proteins or a mixture of both.

5

5.5 MOLECULAR BARCODE USE IN QUALITY CONTROL

In another embodiment, the molecular barcode made up of spike-in tags can be used for quality control. The spike-in tags are added to a sample prior to or during sample processing. The particular profile of spike-in tags added to the sample that make up

10 the unique molecular barcode of the sample are associated (*e.g.*, in a computer database) with information describing the particular sample to which it was added (*e.g.*, name, sample number, container number, origin, process applied, assay performed; or one or more of the foregoing). After sample processing, the actual molecular barcode present in that particular sample is deduced and is associated with the data resulting from the processing of

15 that particular sample. The expected and actual molecular barcodes are compared (schematically represented in Figure 7). The data meet the quality standards only if there is a match between the molecular barcodes. If the data do not meet the quality standards, an error (or non-conformance) report is generated. This serves as notification that the data must be further examined to determine the reason for its failure to meet the quality standard.

20 Data that fail to meet the quality standard are excluded from inclusion in a database containing data from like sample processing. In this way, molecular barcodes can be used to filter out suspect data and prevent the inclusion of such data into a larger database.

In another embodiment, molecular barcodes can be used to identify cross-contamination of samples. If a sample containing a molecular barcode has been cross-

25 contaminated with one or more other samples also containing molecular barcodes, unexpected spike-in tags may be detected along with the expected spike-in tags at the end of sample processing. Spike-in tags can be added at different times during sample preparation and processing to identify the point at which cross-contamination occurred. For example, should only a subset of the contaminating sample's spike-in tags be present in the

30 contaminated sample, then this would indicate the preparation and/or processing step that the contaminating sample had completed at the time of addition.

In another embodiment, molecular barcodes can be used to monitor pool identities. In some instances, it may be desirable to pool multiple samples of different identities at some point during sample preparation or processing. In this way, it can be

35 determined what samples are and are not a member of a certain pool. For example, if the

molecular barcode scheme depicted in Figure 6A is used, eight pools can be made by combining the samples in each of the eight rows. Because there is a spike-in tag common to the members of a row, it can be used as a pool identifier.

5   In a further embodiments, some or all aspects of sample quality control can be done by a computer (see Section 5.6 below).

5.6 IMPLEMENTATION SYSTEMS AND METHODS

The analytical methods of the present invention can preferably be implemented using a computer system, such as the computer system described in this

10   section, according to the following programs and methods. Such a computer system can also preferably store and manipulate a database of the present invention which comprises a compendium of molecular barcodes and which can be used by a computer system in implementing the analytical methods of this invention. Each molecular barcode preferably represents a sample identity profile. Accordingly, such computer systems are also

15   considered part of the present invention. In a specific embodiment, the molecular barcodes stored in digital form in the database (in the compendium) are associated with one or more of the following types of information regarding the biological specimen/sample: name, sample number, container number, origin, process applied, assay performed, and profile of spike-in tag added. By way of example, the molecular barcodes can be associated with such

20   information by way of an index, a relational database, or a sequential search.

In one embodiment, the computer system comprises one or more processing units and one or more memory units connected to said one or more processor units. Said one or more memory units contain one or more programs which cause said one or more processor units to execute steps of:

25   (a)   receiving a first data structure comprising the molecular barcode of a processed sample; and

      (b)   comparing said first data structure to a plurality of data structures in a database, each data structure comprising a molecular barcode associated with sample identity, to identify a molecular barcode that is a match, wherein said

30   match indicates the identity of said sample.

In a further embodiment, the computer system comprises a program that autonomously matches the identities of one or more of the spike-in tags of a processed sample with the compendium of possible molecular barcodes and outputs the identity of said sample.

35

In another embodiment, the computer system performs one or more aspects of the sample quality control. For example, the computer program could read the molecular barcode present in the sample that has been processed and query a database which contains the expected molecular barcode for that sample. Should a match between molecular
5   barcodes not be found, the computer could generate a non-conformance report and refrain from automatically adding the suspect data to the database containing sample possessing data until the quality control issues are further addressed. In a specific embodiment, when sample processing uses a microarray, the computer can read the sample's molecular barcode directly from the raw data represented in a TIFF file of the scanned microarray image.
10   An exemplary computer system suitable for implementing the analytic methods of this invention preferably comprises internal components being linked to external components. The internal components of this computer system include a processor element interconnected with a main memory. For example, the computer system can be an Intel Pentium®-based processor of 200 MHZ or greater clock rate and with 32 MB or more main
15   memory. In a preferred embodiment, the computer system is a cluster of a plurality of computers comprising a head "node" and eight sibling "nodes", with each node having a central processing unit (CPU). In addition, the cluster also comprises at least 128 MB of random access memory (RAM) on the head node and at least 256 MB of RAM on each of the eight sibling nodes. Therefore, the computer systems of the present invention are not
20   limited to those consisting of a single memory unit or a single processor unit.

The external components can include a mass storage. This mass storage can be one or more hard disks that are typically packaged together with the processor and memory. Such hard disk are typically of 1 GB or greater storage capacity and more preferably have at least 6 GB of storage capacity. For example, in a preferred embodiment,
25   described above, wherein a computer system of the invention comprises several nodes, each node can have its own hard drive. The head node preferably has a hard drive with at least 6 GB of storage capacity whereas each sibling node preferably has a hard drive with at least 9 GB of storage capacity. A computer system of the invention can further comprise other mass storage units including, for example, one or more floppy drives, one more CD-ROM
30   drives, one or more DVD drives or one or more DAT drives.

Other external components typically include a user interface device, which is most typically a monitor and a keyboard together with a graphical input device such as a "mouse". The computer system is also typically linked to a network link which can be, *e.g.*, part of a local area network (LAN) to other, local computer systems and/or part of a wide
35   area network (WAN), such as the Internet, that is connected to other, remote computer

systems. For example, in the preferred embodiment, discussed above, wherein the computer system comprises a plurality of nodes, each node is preferably connected to a network, preferably an NFS network, so that the nodes of the computer system communicate with each other and, optionally, with other computer systems by means of the

5   network and can thereby share data and processing tasks with one another.

Loaded into memory during operation of such a computer system are several software components. The software components comprise both software components that are standard in the art and components that are special to the present invention. These software components are typically stored on mass storage such as the hard drive, but can be

10   stored on other computer readable media as well including, for example, one or more floppy disks, one or more CD-ROMs, one or more DVDs or one or more DATs. The software component represents an operating system which is responsible for managing the computer system and its network interconnections. The operating system can be, for example, of the Microsoft Windows™ family such as Windows 95, Window 98, Windows NT or

15   Windows2000. Alternatively, the operating software can be a Macintosh operating system, a UNIX operating system or the LINUX operating system. The software components comprise common languages and functions that are preferably present in the system to assist programs implementing methods specific to the present invention. Languages that can be used to program the analytic methods of the invention include, for example, C and C++,

20   FORTRAN, PERL, HTML, JAVA, and any of the UNIX or LINUX shell command languages such as C shell script language. The methods of the invention can also be programmed or modeled in mathematical software packages that allow symbolic entry of equations and high-level specification of processing, including specific algorithms to be used, thereby freeing a user of the need to procedurally program individual equations and

25   algorithms. Such packages include, *e.g.*, Matlab from Mathworks (Natick, MA), Mathematica from Wolfram Research (Champaign, IL) or S-Plus from MathSoft (Seattle, WA).

The software component comprises analytic methods of the present invention, preferably programmed in a procedural language or symbolic package. For

30   example, the software component preferably includes programs that cause the processor to implement steps of accepting a plurality of sample identity profiles and storing the profiles in the memory. For example, the computer system can accept sample identity profiles that are manually entered by a user (*e.g.*, by means of the user interface). More preferably, however, the programs cause the computer system to retrieve sample identity profiles from

35   a database. Such a database can be stored on a mass storage (*e.g.*, a hard drive) or other

computer readable medium and loaded into the memory of the computer, or the database
can be accessed by the computer system by means of the network.

In one embodiment, the computer readable medium contains an encoded data
structure comprising:

5   (a)   a digital representation of the identity of each sample
                    associated with a molecular barcode;

           (b)   a digital representation of the particular spike-in tags
                    comprising each molecular barcode; and

           (c)   a digital representation of which of said molecular barcodes
10   has been added to which of said samples.

In another embodiment, a sample identity profile contained in a database
and/or loaded into the memory of the computer system is represented by a data structure
comprising a plurality of data fields. In a particular embodiment, the data structure for a
particular sample identity profile comprises separate data fields. A plurality of information
15 can be stored in these data fields about each sample, including but not limited to, one or
more of the following types of information: name, origin, process applied, assay performed,
and profile of spike-in tag added. In another embodiment, one or more of the data fields can
be indexed. The analytic software component comprises programs and/or subroutines
which can cause the processor to perform steps of searching the database for complete
20 information regarding a sample that is identified by only one such parameter.

In addition to the exemplary program structures and computer systems
described herein, other, alternative program structures and computer systems will be readily
apparent to the skilled artisan. Such alternative systems, which do not depart from the
above described computer system and programs structures either in spirit or in scope, are
25 therefore intended to be comprehended within the accompanying claims.

5.7 KITS

The present invention also encompasses kits for use in sample tracking or
quality control during sample processing. In one embodiment, the kit comprises a plurality
30 of separate containers, each container containing one or more spike-in tags such that each
container contains a unique molecular barcode. A label on the outside of each container
could provide an indication of the molecular barcode present inside the container.
Alternatively, the label on each container could provide a container number. In this case, an
informational pamphlet could also be included in the kit which matches container number
35 with the molecular barcode present inside the container.

In another embodiment, kits of the present invention can also include microarrays containing probes capable of detecting all of the possible spike-in tags present in the containers of the kit. In this case, an informational pamphlet could also be included in the kit to show the position of the probes capable of binding spike-in tags on the

5 microarray, and the identity of the spike-in tags (or molecular barcode) thus recognized by each probe.

The above-described components can also be included in diagnostic or forensic kits. For example, diagnostic kits designed to detect single nucleotide polymorphisms (*e.g.*, containing in one or more containers oligonucleotide probes

10 hybridizable to the site of a single polynucleotide polymorphism) or aneuplodies (e.g., containing in one or more containers oligo nucleotide probes hybridizable to various sites on various chromosomes and control DNA originating from a sample known to have a particular aneuploidy) in genomic DNA can incorporate this sample tracking or quality control methodology.

15

The following examples are presented by way of illustration of the present invention, and are not intended to limit the present invention in any way.

6. EXAMPLE 1: Sample Tracking During Microarray Analysis

20

6.1 SPIKE-IN TAG DETECTION WITHIN SAMPLES
6.1.1 TAG CONSTRUCTION AND ISOLATION

The plasmids utilized to synthesize synthetic RNA molecules for use as spike-in tags were constructed with common molecular biological techniques known in the

25 art (Figure 1). First, oligonucleotides that were 60 nucleotides in length were synthesized with the middle 52 nucleotides having a random sequence. These oligonucleotides were cloned into pSP64-E1A plasmids that had been previously digested to completion with XbaI and BamHI. The plasmids were amplified, isolated and sequenced. Those clones containing oligonucleotide inserts were linearized with EcoRI and purified before serving as

30 a template for *in vitro* transcription with SP6 DNA-dependent RNA polymerase. Resulting RNA had a 30 residue poly A tail from the plasmid. Spike-in tag RNA was purified with RNeasy columns before use.

35

6.1.2 cRNA PRODUCTION AND ISOLATION

Twenty four different spike-in tag RNA molecules were mixed together at various concentrations. These collections of probe RNA were spiked into 3 µg of total RNA and the resulting mixture was then used in a reverse transcription reaction (Figure
5   2A). Briefly, buffer, oligonucleotide primer, and deoxyribonucleotides were added to the total RNA. The mixture was heated at 65°C for 5minutes then 37°C for 5 minutes before the addition of 1.0 µl Superscript II reverse transcriptase enzyme. The reverse transcription reaction was performed at 37°C for 50 minutes. 1.0 µl RNAse H was then added to the reaction mixture and incubated at 37°C for 20 minutes. The mixture was then incubated at
10   95°C for 5 minutes before a 3 minute incubation on ice. A second oligonucleotide primer was added to the reaction in the presence of 80 units of Klenow DNA polymerase I and incubated at 37°C for 45 minutes before incubation at 65°C for 10 minutes. Upon the completion of the incubations, the cDNA was purified from the excess oligonucleotide primers and enzymes by binding to a resin. The cDNA was eluted from the resin and
15   prepared for PCR by the addition of buffer, oligonucleotide primers, deoxynucleotides, and 200 units of Taq polymerase. The PCR reaction conditions used were 94°C for 45 seconds, 55°C for 2 minutes, and 65°C for 4 minutes for 10 cycles. After PCR, the product was precipitated with ethanol. The dry pellet was resuspended in 8 µl H$_2$0 and prepared for in vitro transcription by the addition of buffer, deoxyribonucleotides, and T7 polymerase. The
20   mixture was incubated at 37°C overnight. The resulting cRNA was purified before a fluorescent dye was covalently conjugated to the polynucleic acid. The fluorescently labeled cRNA was then purified by resin binding as described previously. The cRNA was fragmented by heating in the presence of zinc.

25   6.1.3 DETECTION ON MICROARRAYS

The resulting labeled cRNA from Section 6.1.2 was then hybridized to an ink jet oligonucleotide microarray (*e.g.*, according to Hughes et al., 2001, *Nature Biotechnology* 19:342-347) displaying DNA probes representing sequences found in public databases as well as probes corresponding to each possible spike-in tag. Figure 2B outlines the
30   procedure as well as the number of tube transfers that can result in handling errors. After hybridization, the microarray was washed and scanned with a fluorescent scanner.
In Figure 3, ten spike-in tags are added to a complex cRNA sample prior to amplification. The same ten spike-in tags are added to a second complex cRNA sample at different concentrations prior to amplification. Each sample is amplified independently and labeled
35   with different fluorescent dyes. The two samples are then combined prior to hybridization

to the microarray. After hybridization (Figure 3A), the ratios were accurately and reproducibly reported. The accuracy is evidenced by the cluster of data points corresponding to each spike-in tag on the plot (Figure 3B). Figure 3 shows that the spike-in tag RNA can be detected on the ink jet oligonucleotide microarray after spike-in tag RNA

5   addition to total sample RNA prior to reverse transcription, cDNA production and isolation, PCR, *in vitro* transcription, cRNA production and isolation, and fluorescent labeling.

Figures 4A-C illustrate microarray images from three experiments in which the same spike-in tag RNA molecules were added to total sample RNA before processing and detection on microarrays. Three different methods of mRNA amplification were used

10   to prepare the biological samples for hybridization to the microarray. In Figure 4A, cDNA was made using an oligo dT primer to reverse transcribe total sample RNA. In Figure 4B, cRNA was made from total sample RNA by reverse transcription before being transcribed *in vitro*. In Figure 4C, cRNA was made from total sample RNA using the protocol outlined in Figure 2A. In all three experiments, spike-in tags were detected on the microarrays.

15

6.2 SPIKE-IN TAGS DO NOT IMPACT DETECTION OF SAMPLE RNA MOLECULES

Figures 5A-C illustrate that spike-in tag RNA molecules do not significantly impact detection of cRNA molecules originating from sample RNA. Two samples of

20   cRNA were prepared, each containing a unique set of spike-in tags and labeled with a unique fluorescent dye. The same complex mixture of cRNA was, however, used in each sample. Both samples were mixed together before hybridization to the microarray. Figures 5A and 5B are fluor reversed pairs (*i.e.*, the experiments are identical except that the fluorescent dyes labeling the cRNA samples have been switched relative to each other).

25   Plots in these figures show the ratio of the intensities detected at each probe on the microarray. The probes corresponding to the spike-in tags are not included in the plots. If the spike-in tags were cross-hybridizing with a probe directed to a gene sequence, there would be significant ratios on the plots. No such ratios were seen. Figure 5C shows the correlation of the ratios in Figures 5A and 5B. If there were any significant ratios detected

30   in either of Figure 5A or 5B, there would be data points lying outside the central cluster. This shows that the spike-in tags do not cross-hybridize with any gene sequence and thus do not interfere with sample hybridization or detection during microarray processing.

35

6.3 SPIKE-IN TAGS CAN FUNCTION AS SAMPLE IDENTITY CONTROLS

Samples in a 96-well plate are identified with spike-in tag RNA molecules coded to the particular position of the sample in the 96-well plate. Because 96-well plates are laid out as eight rows of twelve columns, this is accomplished by the addition of two

5   different spike-in tag RNA molecules into each well. One probe corresponds to the row number while the other corresponds to the column number. Figure 6A illustrates this type of identification scheme.

An alternate identification scheme for molecular barcodes is through the use of sample numbers. A unique spike-in tag is assigned to the numbers 0-9 for each digit of a

10   sample number. Thus ten different spike-in tag RNA molecules are needed for each order of magnitude. In this way, 999 samples can be identified with the creation of just 30 unique spike-in tags. Figure 6B illustrates this type of identification scheme. The upper panel is an example of a key wherein each digit of the sample number is assigned a unique spike-in tag for the numbers 0-9. The lower panel is an example of four different molecular barcodes

15   that can be given to four different samples. In this example, three spike in tags are added to each sample to make the molecular barcode.

6.4 TRACKING OF SAMPLES COLLECTED FROM PATIENTS

After harvesting, the tissue sample from the patient is placed in a tube

20   already containing spike-in tag RNA molecules and processed according to standard protocols. Total RNA is isolated using techniques well-known in the art (*see e.g.*, Ausubel, F. M. et al., eds., 1987-1993, *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc. New York). mRNA is isolated and cRNA is prepared as described above in Section 6.1.2. cRNA is then used to hybridize to a microarray as in Section 6.1.3.

25

7. EXAMPLE 2: Sample Tracking During Other Multi-Step Processes
7.1 SCREENING USING IMMUNOFLUORESCENCE

Tissue culture cells are screened for the presence of a protein of interest through the use of immunofluorescence. In preparation for the assay, cells are directed to

30   express a spike-in tag which is detected simultaneously with the protein of interest. The spike-in tags in this example are synthetic mRNA molecules that can be transcribed by the cell into corresponding amino acids. The amino acids of each spike-in tag provide a unique epitope, none of which are endogenously expressed in the cell. Each epitope provided by each spike-in tag has a corresponding antibody which can specifically recognize the epitope.

35

Briefly, cells are grown on coverslips, fixed, washed, and placed on parafilm on the bottom of a petri dish. Fixed cells are then incubated for one hour at room temperature in a mixture of primary antibodies which binds to the protein of interest as well as the epitopes resulting from the spike-in tags. Cells are then washed and incubated for

5   one hour in a mixture of secondary antibodies that bind to the primary antibodies and are conjugated to fluorescent molecules. After an additional wash, the cell-containing coverslips are mounted on a slide with a drop of gelvatol sealant solution. Presence and location of the protein of interest is visualized by examination of the cells though a fluorescent microscope. The presence of the spike-in tags is also detected in this manner.

10

7.2 LIBRARY CONSTRUCTION

Molecular barcodes are used in library construction in two different ways. First, one or more spike-in tags are added to the sample to be used as an RNA donor for library construction. Total RNA is isolated from sample tissue. mRNA is then further

15   isolated by selecting for the poly A tail-containing subpopulation of RNA. The mRNA is the converted to single-stranded cDNA via reverse transcription using reverse transcriptase. The single-stranded cDNA is converted to double-strand cDNA by the addition of a second primer and DNA polymerase I (*e.g.*, Klenow enzyme). The double-stranded cDNA is then cloned into a bacterial cloning plasmid and amplified. A sub-population of plasmids in the

20   library will have as their insert a nucleic acid sequence corresponding to the sequence of the added spike-in tags. PCR is used to confirm the presence of the expected spike-in tags in the library.

Alternatively, the plasmid to be used in library construction has one or more spike-in tag sequences added its backbone (*i.e.*, at a site other than the multiple cloning

25   site). After library construction as described above (using only donor RNA from the sample as plasmid inserts), every library plasmid member contains the same molecular barcode. This molecular barcode can be distinguished through the use of DNA sequencing. This method of library tagging allows for the distinction of plasmids that originated from different library syntheses.

30

8. REFERENCES CITED

All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by

35   reference in its entirety for all purposes.

Many modifications and variations of the present invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only, and the invention is to be limited only by the terms of the appended claims along with the full scope

5   of equivalents to which such claims are entitled.

10

15

20

25

30

35

What Is Claimed:

        1.   A method for providing internal identification of a sample comprising RNA or protein from a biological specimen such that said sample can be tracked through a processing step, comprising the following steps in the order stated:

        (a)   adding one or more spike-in tags that are RNA or protein molecules that form a first molecular barcode to said sample;

        (b)   processing said sample through a plurality of processing steps;

        (c)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

        (d)   comparing said first molecular barcode with said second molecular barcode to determine whether said first molecular barcode does not differ from said second molecular barcode, wherein a match indicates that said sample has been tracked through said processing step.

        2.   The method of claim 1 wherein the sample comprises RNA.

        3.   The method of claim 1 or 2 wherein the spike-in tags are RNA molecules.

        4.   The method of claim 1 wherein the spike-in tags are protein molecules.

        5.   The method of claim 1 wherein said processing step comprises hybridization of nucleic acid from or derived from said sample to a microarray.

        6.   The method of claim 5 wherein the microarray is used to assay the expression levels of gene transcripts in said biological sample.

        7.   The method of claim 1 wherein said spike-in tags are each an RNA molecule comprising a poly A tail, and said sample comprises mRNA.

        8.   The method of claim 7 wherein said spike-in tags are not known to be, or to encode, any naturally occurring protein.

9.   The method of claim 1 wherein two or more spike-in tags are added to said sample.

10.   The method of claim 1 wherein said processing step comprises a plurality of serial transfers of RNA from said sample, or nucleic acid derived therefrom, to different containers or solid phases.

11.   The method of claim 1 wherein said processing step comprises amplifying said RNA or nucleic acid derived therefrom.

12.   The method of claim 1 wherein said processing step comprises synthesizing cDNA, amplifying cDNA, and producing cRNA.

13.   The method of claim 1, wherein said processing step comprises:

(a)   contacting a positionally-addressable array of polynucleotide probes with RNA molecules from said sample, or nucleic acid derived therefrom, under conditions conducive to hybridization between said probes and said RNA molecules or nucleic acid derived therefrom, wherein said array comprises a plurality of polynucleotide probes of different nucleotide sequences bound to different regions of a support, wherein said plurality of polynucleotide probes comprises (i) probes hybridizable to said spike-in tags, and (ii) probes hybridizable to said RNA molecules or nucleic acid derived therefrom; and

(b)   detecting or measuring hybridization between said polynucleotide probes and (i) said spike-in tags, and (ii) said RNA molecules or nucleic acid derived therefrom.

14.   An RNA molecule comprising:

(a)   a sequence of at least 20 contiguous nucleotides of which at least 50% are random; and

(b)   a poly A tail.

15.   A method of synthesizing an RNA spike-in tag comprising transcribing by use of an RNA polymerase a DNA sequence operably linked to a promoter, wherein said DNA sequence comprises a nucleotide sequence in the range of 20 to 5000 contiguous

nucleotides not known to encode any naturally occurring protein, so as to produce an RNA molecule comprising an RNA sequence encoded by said DNA sequence.

5 16.   An RNA molecule comprising:

(a)   a sequence in the range of 20-5000 contiguous nucleotides of which at least 50% are random; and

(b)   a poly A tail.

17.   The RNA molecule of claim 16 in which the random nucleotides are not

10 known to be a sequence encoded by any naturally occurring genome.

18.   The RNA molecule of claim 16 that does not encode any known naturally occurring protein.

15 19.   An RNA molecule consisting of:

(a)   a sequence not known (i) to be present in any naturally occurring nucleic acid, or (ii) to encode any naturally occurring protein; and

(b)   a poly A tail.

20 20.   A kit comprising a plurality of separate containers, each container containing an RNA molecule of claim 16 or 19, wherein the RNA molecule in each container has a different said sequence.

21.   The kit of claim 20 further comprising a microarray comprising probes that

25 hybridize to said RNA molecules or nucleic acid molecules derived from the RNA molecules.

22.   A method for tracking a plurality of samples, each sample comprising RNA from a different biological specimen, or nucleic acid derived therefrom, comprising the

30 following steps in the order stated:

a)   adding to each sample one or more RNA molecules that are spike-in tags such that the one or more spike-in tags in each sample form a molecular barcode that distinguishes each sample from said other samples;

35

b)   processing each of said samples wherein said processing comprises transferring RNA from each sample, or nucleic acid derived therefrom, to a different container or solid phase;

c)   determining the identity of the spike-in tags in each processed sample to determine the molecular barcode in each processed sample; and

d)   comparing the determined molecular barcodes for at least some of said samples to a compendium of molecular barcodes associated with sample identity to determine the identities of said at least some samples.

23.   The method of claim 22 wherein the plurality of samples is at least 10.

24.   The method of claim 22 wherein the compendium consists of at least 20 molecular barcodes.

25.   The method of claim 22 wherein the compendium is stored on a computer.

26.   The method of claim 22 wherein two spike-in tags are added to each sample.

27.   The method of claim 26 wherein one spike-in tag corresponds to a column and one spike-in tag corresponds to a row such that the position of said sample in a grid is given by said molecular barcode.

28.   The method of claim 22 wherein three spike-in tags are added to each sample.

29.   The method of claim 22 wherein N spike-in tags are added, and each of said spike-in tags represents a different digit in a N-digit identification number represented by said molecular barcode, where N is a whole number equal to or greater than 1.

30.   The method of claim 22 wherein said processing step comprises hybridization of nucleic acid from or derived from said sample to a microarray.

31.   The method of claim 30 wherein the microarray is used to assay the expression levels of gene transcripts in said biological specimen.

32.   The method of claim 22 wherein said spike-in tags are each an RNA molecule comprising a poly A tail, and said sample comprises mRNA.

33.   The method of claim 32 wherein said spike-in tags are not known to encode
5   any naturally occurring protein.

34.   The method of claim 22 wherein said processing step comprises a plurality of serial transfers of RNA from each said sample, or nucleic acid derived therefrom, to different containers or solid phases.
10

35.   The method of claim 22 wherein said processing step comprises amplifying said RNA or nucleic acid derived therefrom.

36.   The method of claim 22 wherein said processing step comprises synthesizing
15 cDNA, amplifying cDNA, and producing cRNA.

37.   A method for tracking a plurality of samples, each comprising RNA from a different biological specimen, or nucleic acid derived therefrom, comprising the following steps in the order stated:
20   a)   adding a molecular barcode to each sample that distinguishes each sample
                from said other samples, said molecular barcode composed of one or more
                RNA molecules that are spike-in tags, said spike-in tags each consisting of
                an RNA molecule comprising a sequence not known (i) to be present in any
                naturally occurring nucleic acid, or (ii) to encode any naturally occurring
25   protein;
        b)   processing each of said samples, wherein said processing comprises
                contacting a positionally-addressable array of polynucleotide probes with
                RNA molecules from said sample, or nucleic acid derived therefrom, under
                conditions conducive to hybridization between said probes and said RNA
30   molecules or nucleic acid derived therefrom, wherein said array comprises a
                plurality of polynucleotide probes of different nucleotide sequences bound to
                different regions of a support, wherein said plurality of polynucleotide
                probes comprises (i) probes hybridizable to said spike-in tags, and (ii) probes
                hybridizable to said RNA molecules or nucleic acid derived therefrom; and
35   detecting or measuring hybridization between said polynucleotide probes and

(i) said spike-in tags, and (ii) said RNA molecules or nucleic acid derived therefrom;

c)   determining the identity of the spike-in tags in each processed sample to determine the molecular barcode in each processed sample; and

d)   comparing the determined molecular barcodes for at least some of said processed samples to a compendium of molecular barcodes associated with sample identity to determine the identities of said at least some samples.

38.   The method of claim 37 wherein the spike-in tags each consist of an RNA molecule comprising:

(a)   a contiguous sequence of at least 20 random nucleotides, said contiguous sequence being greater than 50% of the total nucleotide sequence of said RNA molecule; and

(b)   a poly A tail.

39.   A method of using a computer to determine the identity of a processed sample containing a molecular barcode wherein said molecular barcode comprises one or more spike-in tags, comprising the following computer-implemented steps:

(a)   receiving a first data structure comprising the molecular barcode of said processed sample; and

(b)   comparing said first data structure to a plurality of data structures in a database, each data structure comprising a molecular barcode associated with sample identity, to identify a molecular barcode that is a match, wherein said match indicates the identity of said processed sample.

40.   The method of claim 39 which further comprises the computer-implemented step of excluding data from samples that fail to make said match from inclusion in a database containing data from like sample processing.

41.   The method of claim 39 which further comprises the computer-implemented step of outputting an indication of whether or not a match has been obtained.

42.   The method of claim 39, wherein said molecular barcode is formed by spike-in tags comprising a contiguous sequence of at least 20 nucleotides of which at least 50% of said nucleotides are random.

43.   A method of using a computer to determine the identity of a processed sample containing a molecular barcode wherein said molecular barcode comprises one or more spike-in tags, comprising a computer-implemented step of comparing the molecular barcode in said processed sample to a database containing a plurality of data structures, each

5   data structure comprising a molecular barcode associated with sample identity, to identify a molecular barcode that is a match, wherein said match indicates the identity of said processed sample.

44.   The method of claim 43, wherein said molecular barcode is formed by spike-

10   in tags comprising a contiguous sequence of at least 20 nucleotides of which at least 50% of said nucleotides are random.

45.   A computer system for determining the identity of a sample, said computer system comprising:

15   one or more processor units; and
         one or more memory units connected to said one or more processor units,
     said one or more memory units containing one or more programs which cause said one or more processor units to execute steps of:
         (a)   receiving a first data structure comprising the molecular barcode of a

20   processed sample; and
         (b)   comparing said first data structure to a plurality of data structures in a
                database, each data structure comprising a molecular barcode associated with
                sample identity, to identify a molecular barcode that is a match, wherein said
                match indicates the identity of said processed sample.

25

46.   The computer system of claim 45 wherein said one or more programs further cause said one or more processor units to execute steps of:
         a)   receiving identities of one or more spike-in tags associated with said
                processed sample;

30   b)   deducing the molecular barcode formed by said one or more spike-in tags;
         c)   comparing said first data structure to a plurality of data structures in a
                database, each data structure comprising a molecular barcode associated with
                sample identity, to identify a molecular barcode that is a match, wherein said
                match indicates the identity of said processed sample; and

35   d)   outputting the identity of said processed sample.

47.   The computer system of claim 45, further comprising one or more storage media storing said database.

48.   A computer readable medium containing an encoded data structure, said data
5   structure comprising:

(a)   a digital representation of the identity of each sample associated with a molecular barcode;

(b)   a digital representation of the particular spike-in tags comprising each molecular barcode; and

10   (c)   a digital representation of which of said molecular barcodes has been added to which of said samples.

49.   The computer readable medium of claim 48 wherein said data structures are indexed.

15

50.   A method for providing internal identification of a sample comprising protein from a biological specimen such that said sample can be tracked through a processing step comprising the following steps in the order stated:

(a)   adding one or more spike-in tags that are protein molecules
20   that form a first molecular barcode to said sample;

(b)   processing said sample;

(c)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(d)   comparing said first molecular barcode with said second
25   molecular barcode, wherein a match indicates that said sample has been accurately tracked through said processing step.

51.   A method for providing internal identification of a sample comprising RNA
30 from a biological specimen such that said sample can be tracked through a processing step comprising the following steps in the order stated:

(a)   processing said sample wherein said processing comprises making cDNA from said RNA in said sample, wherein subsequent or during said making one or more spike-in tags that are isolated DNA molecules that form a first
35   molecular barcode are added to said processed sample at a time when the nucleic acids derived from the sample RNA are cDNA, wherein said DNA

molecules each comprise sequences not known (i) to be present in any naturally occurring nucleic acid or (ii) to encode any naturally occurring protein;

(b)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(c)   comparing said first molecular barcode with said second molecular barcode, wherein a match indicates that said sample has been accurately tracked.

52.   A method of providing internal identification of a sample from a biological specimen such that said sample can be tracked through a possessing step comprising the following steps in the order stated:

(a)   adding one or more isolated spike-in tags that form a first molecular barcode to said sample, wherein said spike-in tags are biopolymers; wherein when said biopolymers are DNA molecules, said DNA molecules comprise a sequence not known (i) to be present in any naturally occurring nucleic acid, or (ii) to encode any naturally occurring protein;

(b)   processing said sample through a plurality of processing steps wherein said processing comprises contacting a positionally-addressable array of probes with biopolymers from said sample, or biopolymers derived therefrom, under conditions conducive to interaction between said probes and said sample biopolymer molecules or biopolymers derived therefrom, wherein said array comprises a plurality of probes of different sequences bound to different regions of a support, wherein said plurality of probes comprises (i) probes capable of interacting with said spike-in tags, and (ii) probes capable of interacting with sample biopolymers or biopolymers derived therefrom; and detecting or measuring interaction between said probes and (i) said spike-in tags, and (ii) said sample biopolymers or biopolymers derived therefrom;

(c)   determining the identity of the spike-in tags in said processed sample to determine a second molecular barcode; and

(d)   comparing said first molecular barcode with said second molecular barcode to determine whether said first molecular

barcode does not differ from said second molecular barcode, wherein a match indicates that said sample has been tracked through said processing step.

5

10

15

20

25

30

35

1/12

FIG.1

2/12

1 RNA ——————▶ Pipet RNA samples
                 ——————▶ Add RT cocktail
   2 ss cDNA ——————▶ Add RNAse H
                       ——————▶ Add 2nd Strand cocktail
      3 ds cDNA
                       ——————▶ Phenol extract samples
                          ——————▶ Microcon-30 samples
         4 Purified ds cDNA
                             ——————▶ PCR reaction
            5 Amplified cDNA
                                 ——————▶ Qiaquick purification
               6 Purified PCR product
                                     ——————▶ T7 transcription reaction (IVT)
                  7 Amplified cRNA ——————▶ Rneasy purification
                     8 Purified aa-cRNA ——————▶ Storage at-80°C
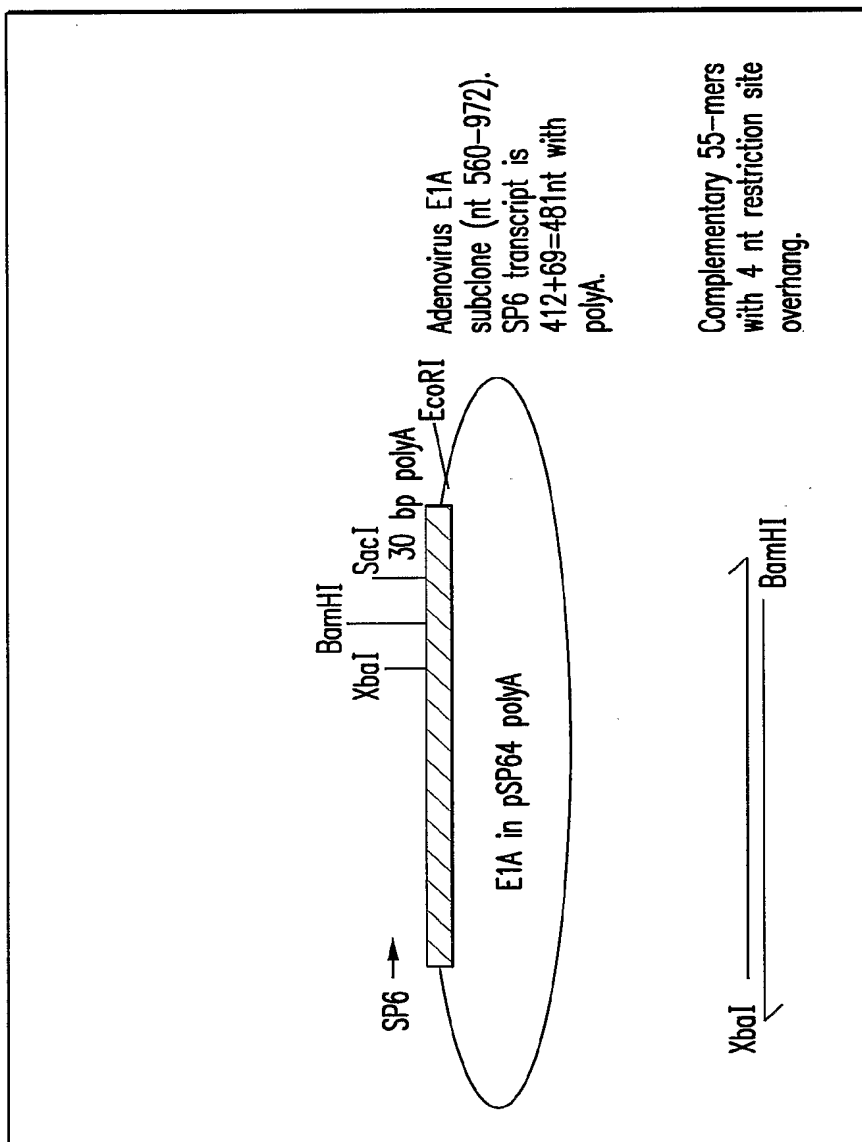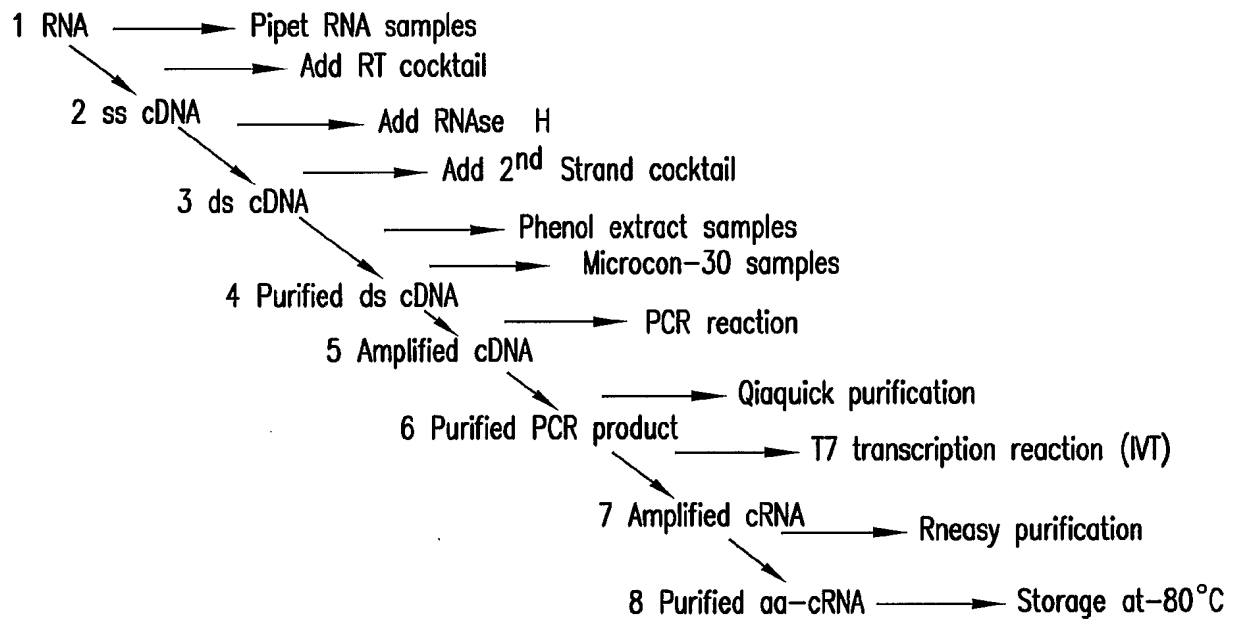
FIG.2A

Method	Transaction
Cell growth/treatment	From plate/flask to harvest tube.
Preparation of total RNA	From harvest tube to final tube
RT–PCR–IVT	Total RNA to RT tube,
	RT tube to phaselock tube
	PI tube to Microcon filter
	Microcon filter to collection tube
	Collection tube to PCR tube/plate
	PCR plate larger tube
	larger tube to Qiaquick column
	Qiaquick column to collection tube
	Collection tube to IVT tube
Purification of cRNA	IVT tube to RNEasy column
	RNEasy column to elution tube
Post–synthetic coupling of cRNA	cRNA tube to new tube
	COMBINE SAMPLES TO BE HYBED
	Coupling tube to Rneasy column
	RNEasy column to elution tube
Setting up the hyb	Elution tube to PCR tube for denaturing sample
	PCR tube to 15 ml conical
	Conical to hyb bag

FIG.2B

4/12
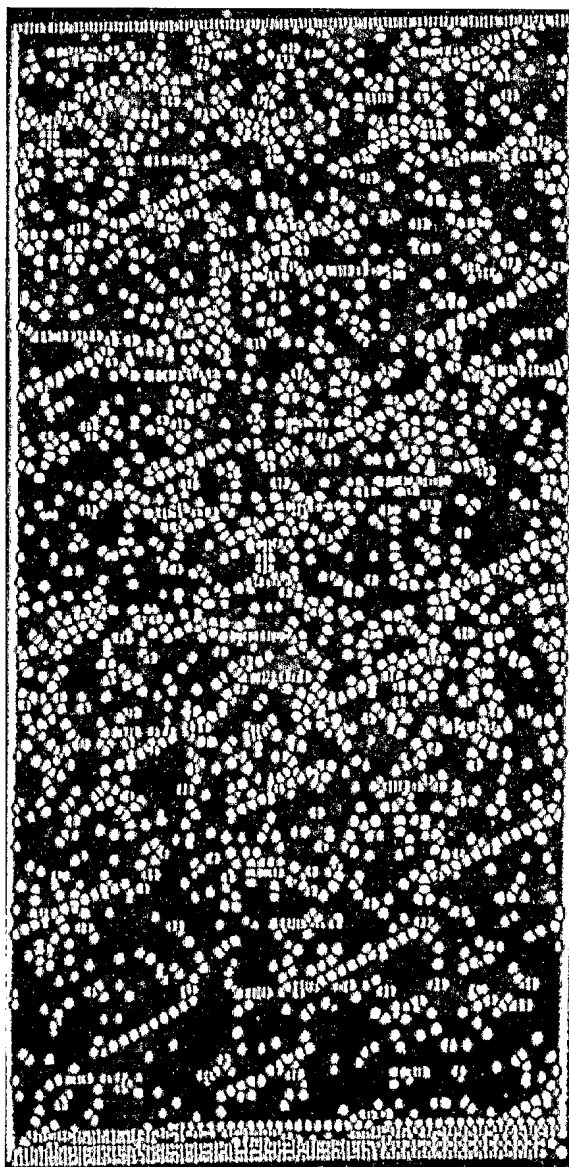

FIG.3A
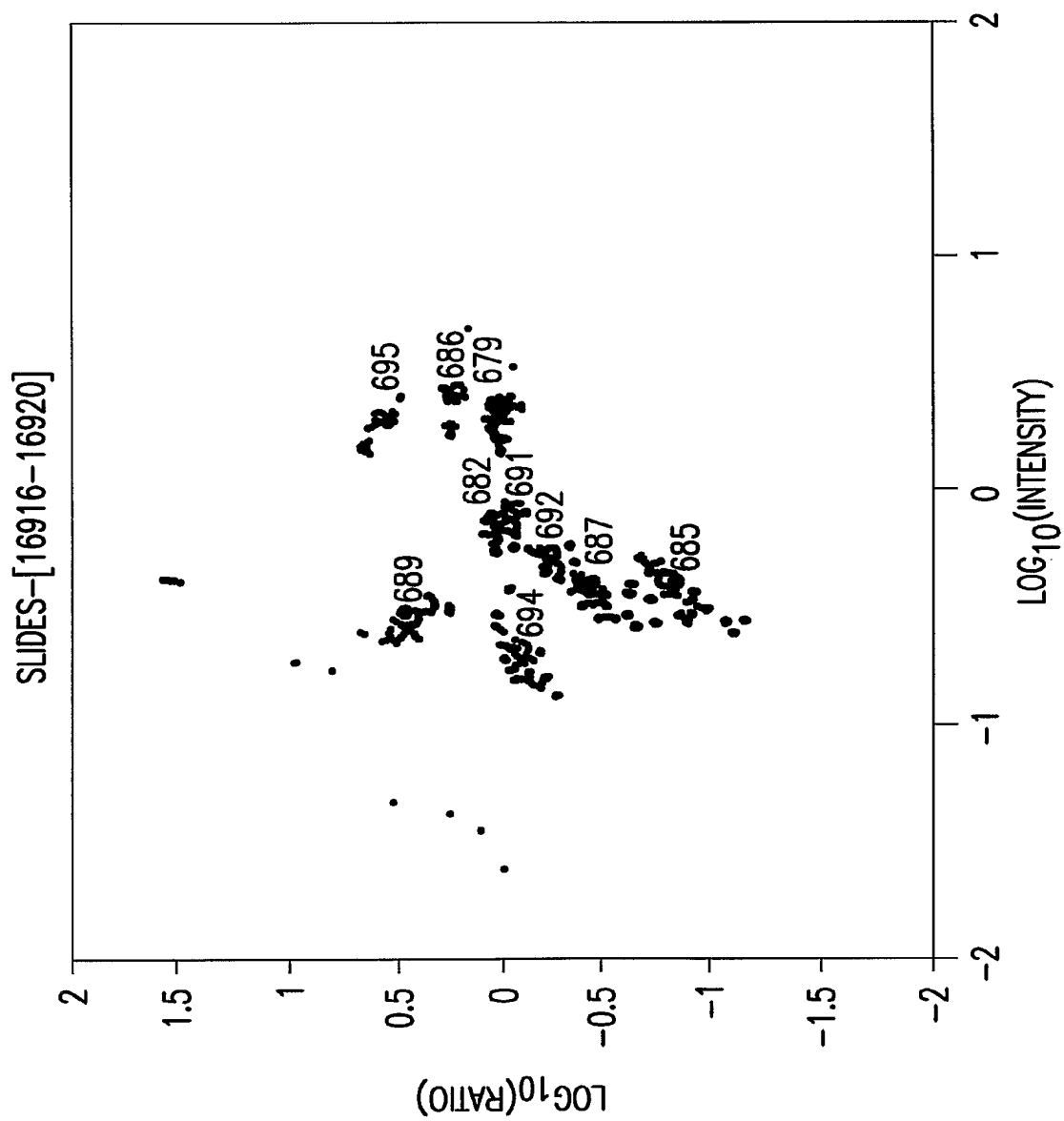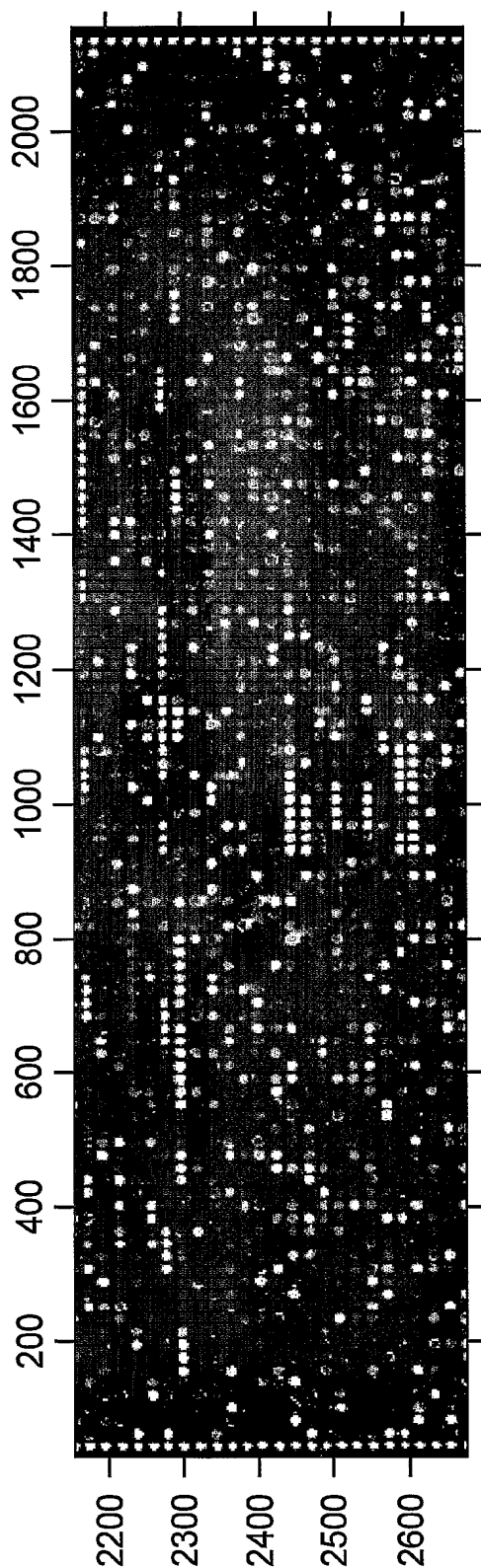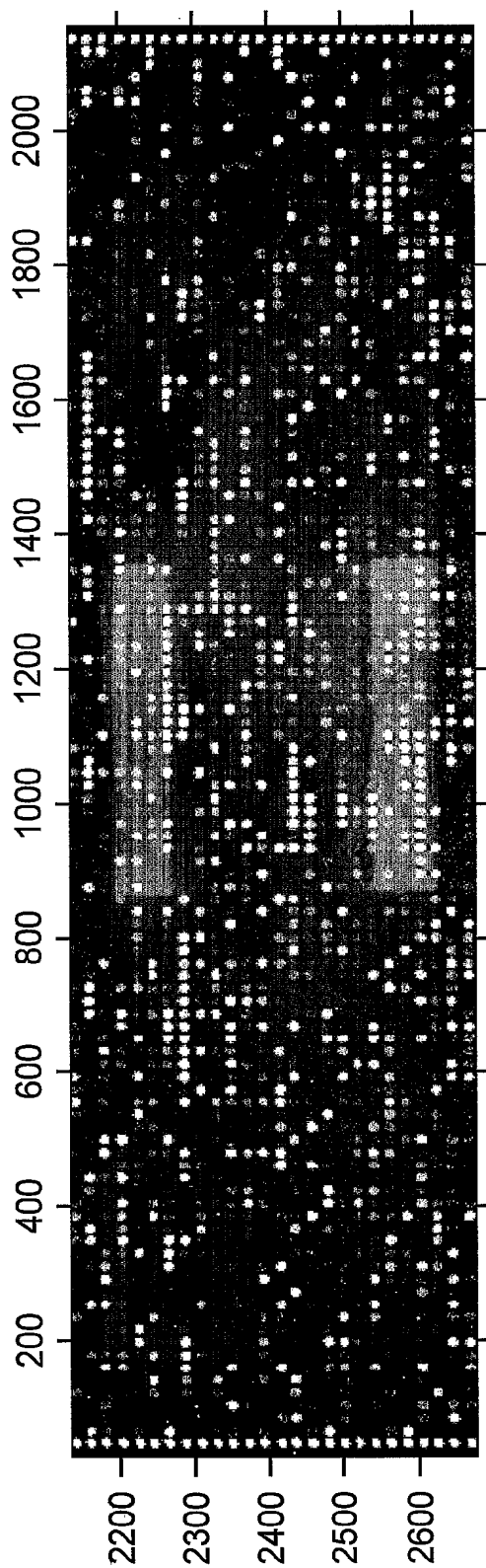
FIG.3B

6/12

FIG.4A

FIG.4B

7/12

FIG.4C

FIG.5A

FIG.5B

FIG.5C

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81	82	83	84
85	86	87	88	89	90	91	92	93	94	95	96

WELL 14

WELL 54

WELL 83
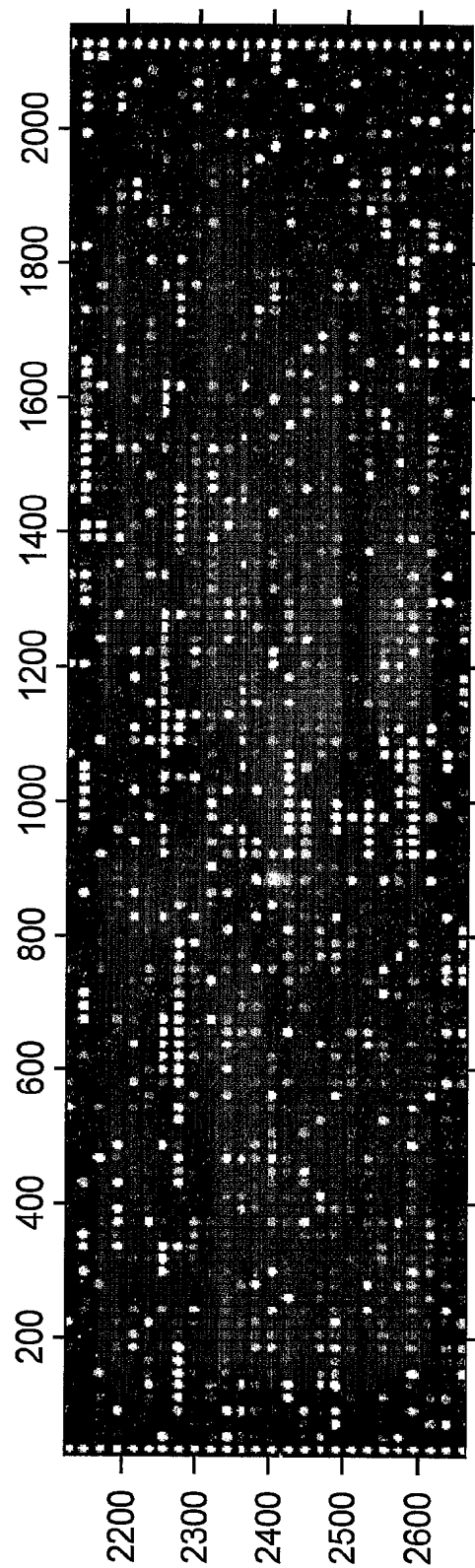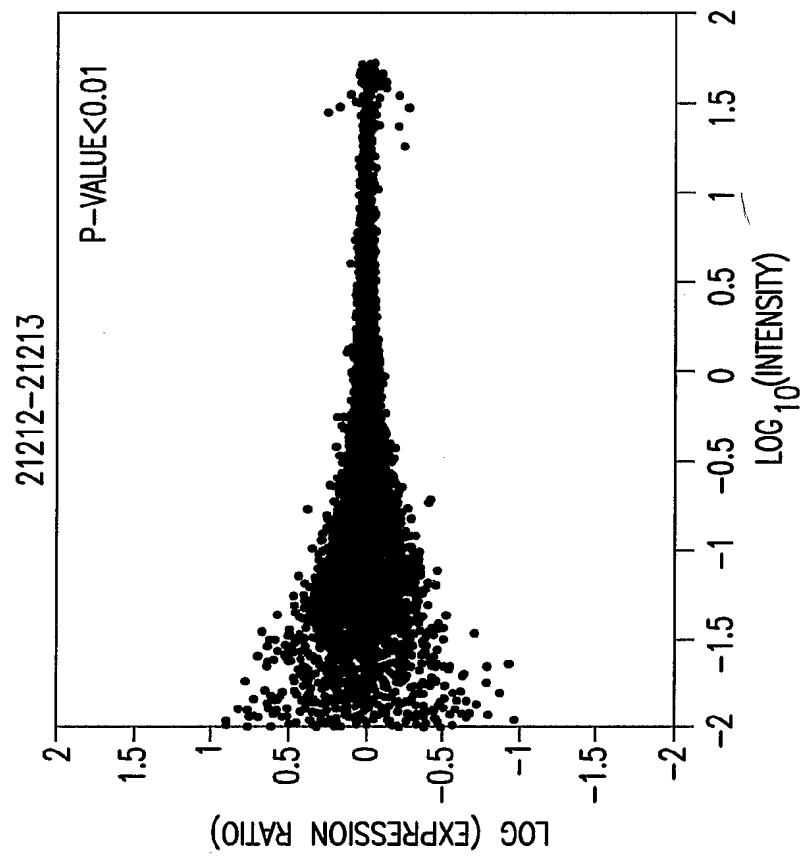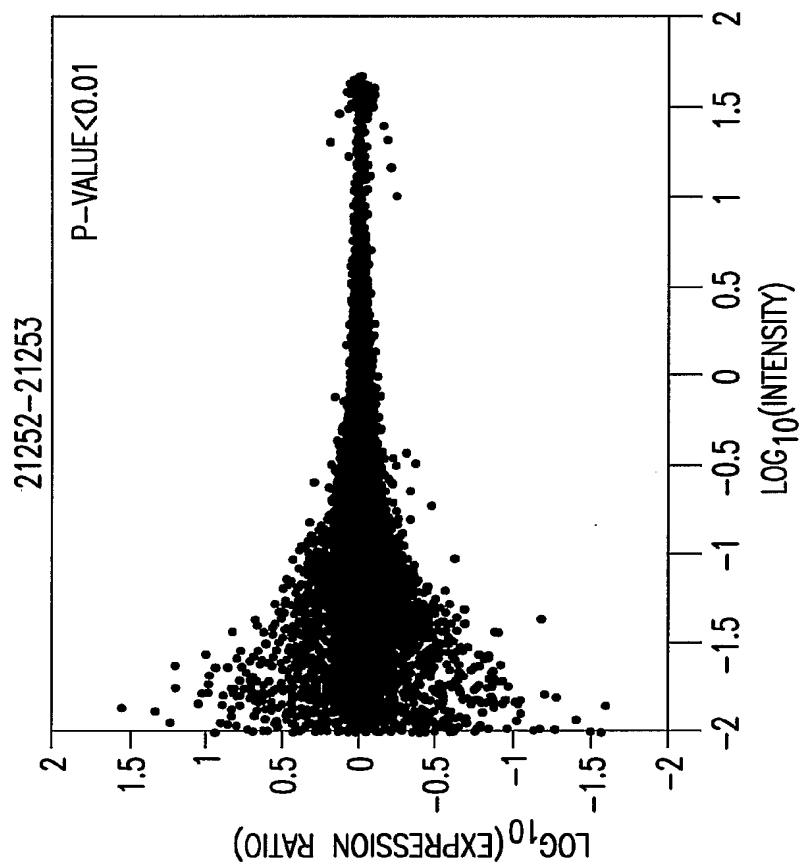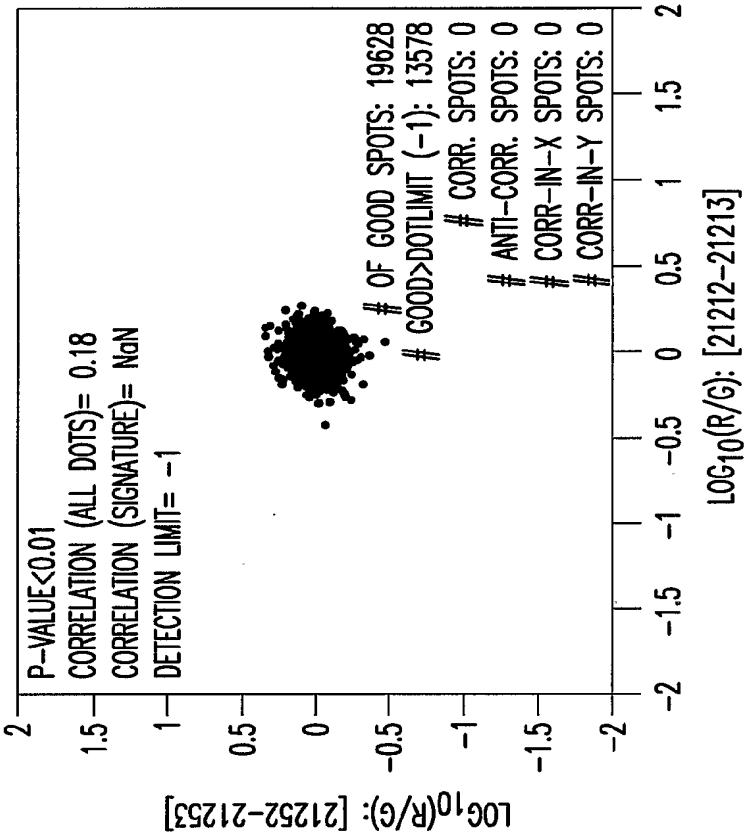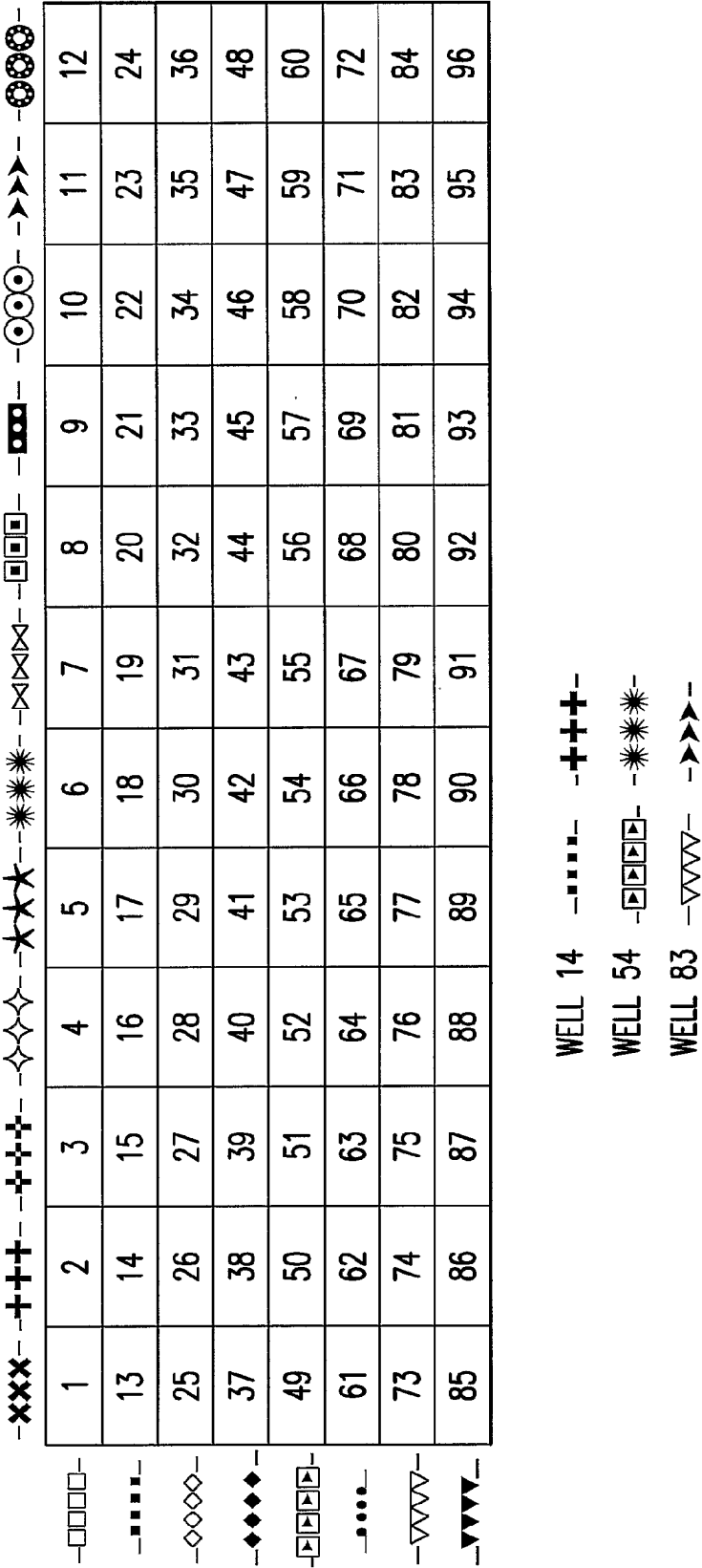
FIG.6A

12/12

	FIRST DIGIT	SECOND DIGIT	THIRD DIGIT
0	—□□□□—	—○ ○○○—	—■·····■—
1	—■■■■—	—●●●●—	—◖○○○○◗—
2	—◇◇◇◇—	—▽▽▽▽—	—✕✕✕✕—
3	—◆◆◆◆—	—▼▼▼▼—	—✚✚✚✚—
4	—▷▷▷▷—	—⋈⋈⋈⋈—	—✚✚✚✚—
5	—▣▣▣▣—	—★★★★—	—◇◇◇◇—
6	—◫◫◫◫—	—⊗⊗⊗⊗—	—✳✳✳✳—
7	—◁◁◁◁—	—◉◉◉◉—	—◉◉◉◉—
8	—▷▷▷▷—	—∧∧∧∧—	—➤➤➤➤—
9	—▣▣▣▣—	—✳✳✳✳—	—⊷⊷⊷⊷—

001   —□□□□—   —□□□□—   —◖○○○○◗—

099   —□□□□—   —✳✳✳✳—   —⊷⊷⊷⊷—

335   —◆◆◆◆—   —▼▼▼▼—   —◇◇◇◇—

999   —▣▣▣▣—   —✳✳✳✳—   —⊷⊷⊷⊷—

FIG.6B

THIS ASSOCIATION IS
MADE AT THE BEGINNING

| SAMPLE INFORMATION, e.g., WHICH SAMPLE WAS PLACED INTO WHICH CONTAINER. | SPIKE—IN TAGS ARE ASSOCIATED WITH A CONTAINER. THE SPIKE—IN TAGS INFER A MOLECULAR BARCODE. | MOLECULAR BARCODE IS READ FROM TIFF IMAGE AND IS ASSOCIATED WITH THE MICROARRAY NUMBER. |

THIS ASSOCIATION IS MADE
AT THE END
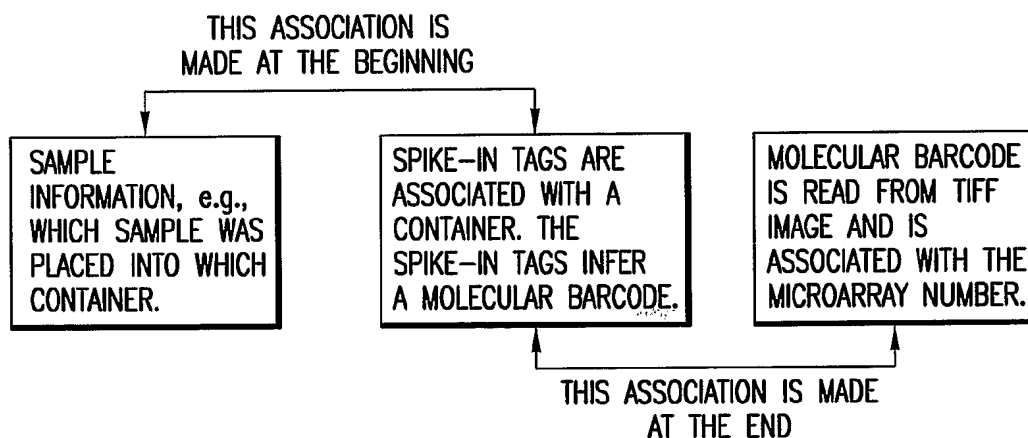
FIG.7

INTERNATIONAL SEARCH REPORT	International application No.
	PCT/US01/48527

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : C12N 15/11; C12Q 1/68; G01N 33/53

US CL : 702/20, 27; 435/6, 7.1; 536/23.1

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 702/20, 27; 435/6, 7.1; 536/23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 00/57183 A1 (BIOVATION LIMITED) 28 September 2000, page 2, line 19-page 3, line 40, page 5, lines 35-45, page 8, lines 28-37, page 9, lines 1-7, page 25, lines 31-35, page 35, lines 2-16.	1-52
A	ARNOT et al. Digital codes from hypervariable tandemly repeated DNA sequences in the Plasmodium falciparum circumsporozoite gene genetically barcode isolates. Molecular and Biochemical Parasitology. 1993, Volume 61, pages 15-24, see the abstract and page 17.	1-52

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

*	Special categories of cited documents:
"A"	document defining the general state of the art which is not considered to be of particular relevance
"E"	earlier application or patent published on or after the international filing date
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O"	document referring to an oral disclosure, use, exhibition or other means
"P"	document published prior to the international filing date but later than the priority date claimed

"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 June 2002 (25.06.2002)	18 JUL 2002
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Marjorie Moran Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998)

Continuation of B. FIELDS SEARCHED Item 3:
EAST, BIOSIS, HCAPLUS, COMPUSCIENCE, BIOTECHDS
search terms: RNA or DNA or nucleic or ?nucleotide; spike or spikein; bar adj code or barcode; protein or peptide or polypeptide